A Computational Perspective of Causal Inference and the Data Fusion Problem

Juan D. Correa

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

A Computational Perspective of Causal Inference and the Data Fusion Problem

Juan D. Correa

The ability to process and reason with causal information is fundamental in many aspects of human cognition and is pervasive in the way we probe reality in many of the empirical sciences. Given the centrality of causality through many aspects of human experience, we expect that the next generation of AI systems will need to represent causal knowledge, combine heterogeneous and biased datasets, and generalize across changing conditions and disparate domains to attain human-like intelligence.

This dissertation investigates a problem in causal inference known as *Data Fusion* [1], which is concerned with inferring causal and statistical relationships from a combination of heterogeneous data collections from different domains, with various experimental conditions, and with nonrandom sampling (sampling selection bias). Despite the general conditions and algorithms developed so far for many aspects of the fusion problem, there are still significant aspects that are not well-understood and have not been studied together, as they appear in many challenging real-world applications.

Specifically, this work advances our understanding of several dimensions of data fusion problems, which include the following capabilities and research questions:

Reasoning with Soft Interventions. How to identify the effect of conditional and stochastic policies in a complex data fusion setting? Specifically, under what conditions can

the effect of a new stochastic policy be evaluated using data from disparate sources and collected under different experimental conditions?

Deciding Statistical Transportability. Under what conditions can statistical relationships (e.g., conditional distributions, classifiers) be extrapolated across disparate domains, where the target is somewhat related but not the same as the source domain where the data was initially collected? How to leverage additional data over a few variables in the target domain to help with the generalization process?

Recovering from Selection Bias. How to determine whether a sample that was preferentially selected can be recovered so as to make a claim about the general underlying super-population? How can additional data over a subset of the variables, but sampled randomly, be used to achieve this goal?

Instead of developing conditions and algorithms for each problem independently, this thesis introduces a computational framework capable of solving those research problems when appearing together. The approach decomposes the query and available heterogeneous distributions into factors with a canonical form. Then, the inference process is reduced to mapping the required factors to those available from the data, and then evaluating the query as a function of the input based on the mapping.

The problems and methods discussed have several applications in the empirical sciences, statistics, machine learning, and artificial intelligence.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First, I would like to thank my mother and grandmother for their unreserved love and support during all my years. They have been, without a doubt, the main necessary causes of all my achievements. Together with them, I have been very fortunate to receive the support of a very loving family. Thanks to my sister, all my aunts, uncles, and cousins for accompanying me from a distance.

I also have to thank my dear friends Adriana, Luz Angela, and Jorge for their friendship, their encouragement, and their support, especially during the years of being a Ph.D. student.

I also want to thank the members of my thesis defense committee: Professors David Blei, Augustin Chaintreau, Judea Pearl, and Jin Tian who have been very kind and accommodating. Special thanks to David and Jin for their insight and encouragement during the thesis proposal and candidacy exam.

I am also indebted to the people I have been fortunate to work with at the Causal AI lab: Amin Jaber, Chris Jeong, Yonghan Jung, Daniel Kumor, Sanghack Lee, Adèle Ribeiro, Kevin Xia, and Justin Zhang. They have all been great colleagues and friends.

Finally, I want to thank my advisor, Professor Elias Bareinboim. During these years as his student, he has kindly shared his passion for science and kept high expectations in my work. I feel fortunate to have had the chance to do research under his guidance.

Chapter 1: Introduction

The success of artificial intelligence and machine learning in several tasks involving high-dimensional predictions had led to growing expectations regarding their autonomy and capability of exhibiting human-level intelligence [2, 3]. These expectations met with fundamental challenges, in particular, due to the absence of proper and explicit treatment of causal inference capabilities [4, 5]. Understanding cause-and-effect relations is one of the pillars of modern science and particularly critical for building AI systems capable of supporting scientific discovery and decision-making under uncertainty.

Controlled experimentation [6] is one of the most pervasive methods to probe for such relations, deemed the "gold standard" for scientific research in many empirical circles. Direct experimentation, however, is oftentimes infeasible due to financial, technical, or ethical considerations. Conditions under which a causal effect can be computed from observational studies, which are generally easier to conduct, have been extensively studied in the literature. Several conditions have been proposed to solve this problem [7, 8, 9], as well as systematic approaches based on graphical models [10, 11, 12, 13, 14].

Even when experimentation can be performed in a specific environment (e.g., in reinforcement learning), another major challenge arises regarding the generalizability of such effects to new settings. AI systems are expected to generalize to new circumstances that they have not been specifically programmed or trained for. For tasks involving observational predictions, various efforts have been made on "transfer learning" [15], "domain adaptation" [16, 17, 18], and "dataset shift" [19] problems. The challenge of generalizing interventional effects to new domains is pervasive throughout the empirical sciences. Even under ideal conditions where these effects can be established through direct experimentation, most findings need to be generalized to a broader, or even different, *target*

domain (e.g., population, setting, environment). In medicine, for example, many drugs are developed and tested using rats in a laboratory, while the intent is almost invariably to understand how humans would react to the treatment. If the conclusions obtained from a study or training process can be extrapolated to the target domain, the process is said to have *external validity*. Constructing studies with external validity is considered one of the main research challenges by the current generation of data scientists [20].

These data fusion challenges are orthogonal and could appear together when the task entails the generalization of causal relationships using data collected from different domains and under different experimental and sampling conditions. For illustration, we consider the following example in the context of medical evaluation.

Example 1 (Data Fusion Challenge)**.** Greenhouse, Kaizar, Kelleher, Seltman, and Gardner [21] discussed the challenges of generalizing studies on the risk of suicidality among pediatric antidepressant users. On investigating the causal relationship between antidepressant use and the risk of suicide attempts, the FDA performed several RCTs, finding that youths receiving antidepressants (X) had approximately twice the number of suicidal thoughts and behaviors (Y) compared to the control groups. These results led to a new policy and the addition of a strict warning on the drugs' label.

Surprisingly, following the warning, reports suggested a decrease in the number of prescriptions and increasing suicidal events in the corresponding age groups. Furthermore, several observational studies found a decreased risk of suicide in patients treated with the same antidepressants, even after adjusting for access to mental health care and other confounding factors.

Some of the possible explanations for this discrepancy are:

- Different populations: There is a mismatch between the study population and the general clinical population regarding ethnicity, race, and income (covariates named E).
- Non-random sampling: By excluding youths with elevated baseline risk for suicide (B)

from their cohorts, based on the inclusion-exclusion criteria, FDA's studies sampled from a distinct population.

The two issues in this example suggest that the internal validity that is achieved following the proper randomization of the treatment assignment is not enough to ensure the generalizability of its findings. The challenge of external validity stems from various reasons, including the data being collected from different populations, under different experimental conditions, and non-random sampling conditions.

In the context of causal inference, the problem of inferring observational and interventional quantities from a combination of heterogeneous datasets is known as the *Data fusion* problem [22]. Previous research on data fusion has focused on when the interventions are atomic, as described by the do-operator [13, 23, 24, 25, 26, 27]. However, there is still virtually no understanding of the conditions under which the effects of stochastic interventions can be generalized across settings. Moreover, it is poorly understood how to reason about interventions when multiple data fusion dimensions appear together, including when confounding and selection biases are both present.

More specifically, our goal in this dissertation is to solve these problems and advance the science of causal inference and data fusion, and answer the following research challenges:

I. Generalizability of the Effect of Soft Interventions. To predict the effect of a (conditional or stochastic) policy in a domain of interest leveraging observational and experimental studies from other domains (environments, settings, populations).

For instance, consider a rover trained in the Californian desert for digging rocks. After exhaustive months of training, NASA wants to deploy the vehicle on Mars, where the environment is somewhat similar, but not the same as on Earth. How to program the rover to operate effectively on Mars with minimal "experimentation" (i.e., trial-and-error) by leveraging the causal knowledge acquired on Earth?

II. Statistical Transportability. To determine whether a conditional probability distribution $P(y \mid x)$ in a target domain is estimable from observational data collected in a training domain while using a minimal amount of data from the target domain.

For example, can we predict the likelihood of a customer buying a product in a particular online store based on customers' shopping behavior on another platform, say, in a physical store?

III. Selection Bias. To decide whether a quantity computed from a subpopulation sample can be recovered and used to make a claim about the entire population when it is known that the selection process of the subpopulation is not random, and not representative of the underlying population.

For instance, can we evaluate the impact of a training program on salary, given that most of the surveyed participants work in the finance sector?

IV. Combined Data Fusion. How can the challenges of transportability, selection bias, limited observability, and evaluating soft interventions be addressed together?

For instance, as discussed in example 1, how to predict the effect of a new protocol for the assignment of antidepressants based on the combination of observational and experimental studies collected in a selected part of the population?

1.1 Outline of the Dissertation

After all, the discussion about the different data fusion challenges tackled in this dissertation will be organized as follows. In chapter 3, we define an algorithmic framework to solve classic causal inference tasks in a systematic and unified fashion, including identifiability of causal effects from observational data [10, 12, 11], identifiability from surrogate experiments [13, 14], and transportability across changing conditions [28, 23, 29, 25, 30].

Chapter 4 discusses the identification of the effect of conditional and stochastic policies and introduces a calculus, akin to do-calculus [31], for reasoning about the effect of these types of interventions (section 4.2). Moreover, section 4.3 extends the proposed algorithmic framework to the task of generalizing the effect of policies across multiple domains with a combination of observational and experimental data.

Chapter 5 considers the problem of generalizing observational distributions (i.e., $P(\mathbf{y} \mid \mathbf{x})$ — also called statistical transportability, dataset shift or domain adaptation. Some sufficient graphical conditions are introduced in section 5.1. Sections 5.2 and 5.3, extend our framework to deal with input distributions where only a subset of the variables is observed and the queries are conditional distributions, respectively. Section 5.4 provides a sufficient and necessary algorithmic approach for this task.

Chapter 6 discusses the problem of identification from data with selection bias (non-random sampling). Specifically, section 6.2 introduces a complete adjustment criterion for recovering causal effects from a combination of selection biased data and unbiased measurements of some covariates. It also introduces an adjustment criterion to generalize selection-biased experimental findings (i.e., results from a randomized controlled experiment) to a different domain (section 6.3). Then, section 6.4 develops a complete algorithm for recovering from selection-biased data alone which is also sufficient for the case when additional unbiased data is available.

Chapter 7 summarizes the tasks addressed in the previous chapters (section 7.1), the operators developed to solve them (section 7.2), and the algorithmic framework developed throughout the dissertation (section 7.3). Finally, section 7.4 discusses the solution of general data fusion tasks combining the different dimensions studied along chapters 4 to 6 through the framework and gives final remarks. Section 7.5 discusses future work.

1.2 Contributions

This dissertation makes contributions to several problems of causal inference and data fusion that are often studied separately. The new results are presented within an overarching *algorithmic framework* (definition 7, algorithm 1), which will also encompass several canonical tasks such as identification of causal effects from observational data, identification from an arbitrary combination of observational and experimental datasets, and generalization of causal effects from multiple domains and experimental conditions.

In terms of the main technical contributions, this dissertation,

In Chapter 4,

- introduces a new set of inference rules to reason about the effect of soft interventions (theorem 5), which is called *-calculus*, which appeared in

  B J. D. Correa and E. Bareinboim, "A Calculus For Stochastic Interventions: Causal Effect Identification and Surrogate Experiments," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020;

- derives a complete graphical and algorithmic characterization (necessary and sufficient) for the *transportability (generalization) of soft interventions* from arbitrary experimental distributions and multiple domains (theorem 6), which appeared in

  B J. D. Correa and E. Bareinboim, "General Transportability of Soft Interventions: Completeness Results," in *Advances in Neural Information Processing Systems*, 2020;

In Chapter 5,

- develops graphical conditions and a complete algorithm for the *generalization of observational (conditional) distributions across domains* with limited data in the target domain (theorem 8), which appeared in

  B J. D. Correa and E. Bareinboim, "From Statistical Transportability to Estimating the Effects of Stochastic Interventions," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2019;

In Chapter 6,

- introduces a complete adjustment criterion for the *recoverability of causal effects* from selection bias and partial unbiased data (definition 22, theorem 9), which appeared in

  B J. D. Correa and E. Bareinboim, "Causal Effect Identification by Adjustment under Confounding and Selection Biases," in *Proceedings of the Thirty-First Conference on Artificial Intelligence*, 2017 and

  B J. D. Correa, J. Tian, and E. Bareinboim, "Generalized Adjustment Under Confounding and Selection Biases," in *Proceedings of the 32th Conference on Artificial Intelligence*, 2018;

- develops complete (necessary and sufficient) algorithmic and graphical conditions for the *recoverability of causal effects* from selection biased data (theorem 18), which appeared in

  B J. D. Correa, J. Tian, and E. Bareinboim, "Identification of Causal Effects in the Presence of Selection Bias," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019;

- develops a sound algorithm for the *recoverability of causal effects* from a combination of selection biased data *and nonbiased partial data* (theorem 20), which appeared in the same paper as the previous contribution; and

- introduces a complete (necessary and sufficient) adjustment criterion for the *generalization of selection-biased experimental studies* across domains (definition 27, theorems 13 and 15), which appeared in

  B J. D. Correa, J. Tian, and E. Bareinboim, "Adjustment Criteria for Generalizing Experimental Findings," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

In Chapter 7,

- proves the soundness and completeness of the proposed algorithm for transportability of conditional and stochastic policies (theorem 21).

- proves the soundness of the algorithm for arbitrary data fusion tasks (theorem 22).

Chapter 2: Notation and Logical Foundations

This chapter introduces the notation used throughout the dissertation, as well as the Structural Causal Model (SCM) framework [31, sec. 7.1], [39], which serves as the logical foundation for the analysis of the problems discussed in the following chapters.

2.1 Notation

Random variables are denoted by capital letters, X, and their values by small letters, x. Bold upper case and lower case letters, such as $\mathbf{X}$ and $\mathbf{x}$, represent sets of variables and sets of values, respectively. The domain of a variable X is denoted by $\text{Val}(X)$, while the domain of a set of variables is defined as $\text{Val}(\mathbf{X}) = \text{Val}(X_1) \quad \text{Val}(X_2) \quad$, for $X_i \in \mathbf{X}$. Two sets of values $\mathbf{x}$ and $\mathbf{z}$ are said to be consistent if they assign the same value for every $\mathbf{X} \setminus \mathbf{Z}$. Moreover, $\mathbf{x}(\mathbf{Z})$ denotes the subset of values in $\mathbf{x}$ corresponding to variables in $\mathbf{Z}$.

Throughout this document, directed acyclic graphs (with bidirected arrows) are heavily used to encode assumptions. They will be represented with calligraphic letters, such as $\mathcal{G}$, or $\mathcal{H}$. Moreover, $\mathcal{G}_{\overline{\mathbf{W}}\underline{\mathbf{X}}}$ denotes the subgraph resulting from removing edges coming into vertices $\mathbf{W}$ and going out from vertices $\mathbf{X}$. Also, $\mathcal{G}[\mathbf{C}]$ denotes the subgraph of $\mathcal{G}$ made only of nodes in $\mathbf{C} \quad \mathbf{V}$ and the edges between them.

Common kinship relationships such as parents, descendants, and ancestors of a variable (or a set of variables) are often used. For example, the set of parents of $\mathbf{X}$ in $\mathcal{G}$ is defined as $Pa(\mathbf{X})_{\mathcal{G}} := \mathbf{X} \cup \bigcup_{X \in \mathbf{X}} Pa(X)_{\mathcal{G}}$. The other relationships are denoted by $De()$ and $An()$ are defined analogously. Note those sets always include X itself. The subscript $\mathcal{G}$ is often omitted when the graph is clear from the context.

2.2 Structural Causal Models and Interventions

The language of *Structural Causal Models* (SCMs) can be used to describe the collection of mechanisms of a phenomenon of interest. SCMs provide a flexible formalism for data-generating processes and can accommodate several approaches to causal inference in the literature [40, 41, 42, 43].

Definition 1 (Structural Causal Model (SCM) [44, 39])**.** A structural causal model M is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) \rangle$, where

- $\mathbf{U}$ is a set of background variables, also called exogenous variables, that are determined by factors outside the model;
- $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called endogenous, that are determined by other variables in the model – that is, variables in $\mathbf{U} \cup \mathbf{V}$;
- F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $\mathbf{U}_i \cup \mathbf{Pa}_i$ to V_i, where $\mathbf{U}_i \subseteq \mathbf{U}, \mathbf{Pa}_i \subseteq \mathbf{V} \setminus \{V_i\}$, and the entire set F maps $\mathbf{U}$ to $\mathbf{V}$. That is, for $i = 1, \ldots, n$, each $f_i \in F$ is such that

$$v_i \leftarrow f_i(\mathbf{pa}_i, \mathbf{u}_i), \tag{2.1}$$

  that is, it assigns a value to V_i depending on (the values of) a select set of variables in $\mathbf{U} \cup \mathbf{V}$; and
- $P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$.

Throughout this dissertation, it is assumed that there are no cyclic dependencies among the functions in F; as a result, F induces a unique solution for $\mathbf{V}$ given any value $\mathbf{u} \in \mathbf{U}$.[1]

Interventions are represented by changes in the mechanism of one or more variables in an SCM. An intervention $do(\mathbf{X} = \mathbf{x})$ entails the replacement of the original mechanisms of each $X \in \mathbf{X}$ for constant values $x \in \mathbf{x}$. Such an intervention produces a new SCM, defined as follows.

[1] Such models are called *recursive* SCMs.

Definition 2 (Submodel – "Interventional SCM"). Let $M = \langle \mathbf{U}; \mathbf{V}; F; P(\mathbf{U}) \rangle$ be a causal model, $\mathbf{X}$ a set of variables in $\mathbf{V}$, and $\mathbf{x}$ a particular realization of $\mathbf{X}$. A submodel $M_{\mathbf{x}}$ of M is the causal model

$$M_{\mathbf{x}} = \langle \mathbf{U}; \mathbf{V}; F_{\mathbf{x}}; P(\mathbf{U}) \rangle; \tag{2.2}$$

where

$$F_{\mathbf{x}} = \{ f_i : V_i \notin \mathbf{X} \} \cup \{ \mathbf{X} = \mathbf{x} \}. \tag{2.3}$$

To reason about the behavior of the observable variables in the original model M or a submodel $M_{\mathbf{x}}$, we can look at each particular configuration $\mathbf{u}$ of the exogenous variables and the responses they induce in the observable variables through the mapping F. Those responses are formally defined as follows.

Definition 3 (Potential Response). Let $\mathbf{X}$ and $\mathbf{Y}$ be two sets of variables in $\mathbf{V}$, and $\mathbf{u}$ be a unit. The potential response $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is defined as the solution for $\mathbf{Y}$ of the set of equations $F_{\mathbf{x}}$ with respect to SCM M (for short, $\mathbf{Y}_{M_{\mathbf{x}}}(\mathbf{u})$). That is, $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{Y}_{M_{\mathbf{x}}}(\mathbf{u})$.

While potential responses are defined for a particular configuration $\mathbf{u}$ of the unobserved variables, averaging over all $\mathbf{u}$ gives rise to qualitatively different types of distributions. These distributions are classified in layers according to the Pearl Causal Hierarchy (PCH) [4, 39].

Definition 4 (PCH Valuation [39, 31]). An SCM $M = \langle \mathbf{U}; \mathbf{V}; F; P(\mathbf{U}) \rangle$ induces probability distributions for each of the three layers (L_1, L_2 and L_3) of the hierarchy. For $\mathbf{Y}; \mathbf{X}; \mathbf{Z}; \mathbf{W}$ $\mathbf{V}$:

$$L_1 \text{ (associational)}: \quad P^M(\mathbf{y}) = \sum_{\{\mathbf{u} \mid \mathbf{Y}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u}); \tag{2.4}$$

$$L_2 \text{ (interventional)}: \quad P^M(\mathbf{y_x}) = \sum_{\{\mathbf{u} \mid \mathbf{Y_x}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u}); \tag{2.5}$$

$$L_3 \text{ (counterfactual)}: \quad P^M(\mathbf{y_x}, \ldots, \mathbf{z_w}) = \sum_{\{\mathbf{u} \mid \mathbf{Y_x}(\mathbf{u})=\mathbf{y}, \ldots, \mathbf{Z_w}(\mathbf{u})=\mathbf{z}\}} P(\mathbf{u}). \tag{2.6}$$

Several results in this dissertation pertain to the identification of interventional quantities (L_2), usually called *causal effects*.

Definition 5 (Causal effect [31, Def. 3.2.1])**.** Let $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be two disjoint sets of observable variables. The causal effect of $\mathbf{X}$ on $\mathbf{Y}$, denoted $P(\mathbf{y_x})$ or $P(\mathbf{y} \mid do(\mathbf{x}))$ is a function from $\mathbf{X}$ to the space of probability distributions on $\mathbf{Y}$. For each $\mathbf{x} \in \text{Val}(\mathbf{X})$, $P(\mathbf{y} \mid do(\mathbf{x}))$ represents the probability of $\mathbf{Y} = \mathbf{y}$ in the submodel $M_{\mathbf{x}}$.

Any approach to causal inference requires some sort of causal assumption about the data generating process [39]. This is formally known as the *Causal Hierarchy Theorem* (CHT). The CHT says that knowledge from L_1 underdetermines L_2, and L_1, L_2 underdetermine L_3. Causal diagrams encode partial knowledge about the SCM that allows for principled and nonparametric causal inference [39, sec. 1.3], defined next.

Definition 6 (Causal Diagram [39])**.** Consider an SCM $M = \langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) \rangle$. Then, G is said to be a *causal diagram* (of M) if constructed as follows:

(1) add a vertex for every endogenous variable in the set $\mathbf{V}$,

(2) add an edge ($V_j \rightarrow V_i$) for every $V_i, V_j \in \mathbf{V}$ if V_j appears as an argument of $f_i \in F$.

(3) add a bidirected edge ($V_j \dashleftarrow\dashrightarrow V_i$) for every $V_i, V_j \in \mathbf{V}$ if the corresponding $\mathbf{U}_i, \mathbf{U}_j \subseteq \mathbf{U}$ are not jointly independent, or the corresponding functions f_i, f_j share some $U \in \mathbf{U}$ as an argument.

The structural information captured by the causal diagram —the functional dependency of each observable on other endogenous and exogenous variables— leads to a factorization over observable families (instead of $P(\mathbf{U})$ as in eq. (2.4)) for expressing $P(\mathbf{V})$ as a product with one factor per observable. To understand how this works, first notice

$$P(\mathbf{v}) = \sum_{\{\mathbf{u} | \mathbf{V}(\mathbf{u}) = \mathbf{v}\}} P(\mathbf{u}) \tag{2.7}$$

$$= \sum_{\mathbf{u}} P(\mathbf{V}(\mathbf{u}) = \mathbf{v}) P(\mathbf{u}); \tag{2.8}$$

where $P(\mathbf{V}(\mathbf{u}) = \mathbf{v})$ is either 0 or 1 since once $\mathbf{u}$ is fixed, the potential response $\mathbf{V}(\mathbf{u})$ is deterministic and either equal to $\mathbf{v}$, or not. Moreover, the probability $P(\mathbf{V}(\mathbf{u}) = \mathbf{v})$ can be decomposed as

$$\prod_{V_i \in \mathbf{V}} P(V_i(\mathbf{u}) = \mathbf{v}(V_i)). \tag{2.9}$$

Individually, for each $V_i \in \mathbf{V}$, $V_i(\mathbf{u}) = f_i(\mathbf{pa}_i(\mathbf{u}), \mathbf{u}_i)$, hence

$$(V_i(\mathbf{u}) = \mathbf{v}(V_i)) = (V_i = \mathbf{v}(V_i) \mid \mathbf{pa}_i, \mathbf{u}_i) \tag{2.10}$$

Then,

$$P(\mathbf{V}(\mathbf{u}) = \mathbf{v}) = \prod_{V_i \in \mathbf{V}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i); \tag{2.11}$$

where $v_i = \mathbf{v}(V_i)$, $\mathbf{pa}_i = \mathbf{v}(\mathbf{Pa}_i)$ and $\mathbf{u}_i = \mathbf{u}(\mathbf{U}_i)$. Indeed, $\{\mathbf{u} \mid \mathbf{V}(\mathbf{u}) = \mathbf{v}\}$ in eq. (2.7) is exactly the same set for which eq. (2.11) is 1 and not 0. Consequently,

$$P(\mathbf{v}) = \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}); \tag{2.12}$$

as we are effectively summing $P(\mathbf{u})$ for the same sets $\mathbf{u}$. We call eq. (2.12) the *latent factorization* of $P(\mathbf{V})$.

A similar argument could be used to write $P(\mathbf{v_x}) = P(\mathbf{v} \mid do(\mathbf{x}))$ in terms of factors of

the form $P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i)$. There is, however, another approach that leverages the already derived eq. (2.12). Consider any model M and an intervention $do(\mathbf{x})$. Let $M' = M_{\mathbf{x}}$ and let P' denote the observational distribution of M'. The causal diagram induced by M' is $G' = G_{\overline{\mathbf{X}}}$. Then,

$$P'(\mathbf{v}) = \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V}} P'(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}). \tag{2.13}$$

Here $M_{\mathbf{x}}$ models a system where the variables in $\mathbf{X}$ are held constant with values $\mathbf{x}$ and have no parents (i.e., $\mathbf{Pa}^{\prime}_{\mathbf{x}} = \emptyset$). Then, the observational distribution P' is the same as the interventional distribution $P(\mathbf{v} \mid do(\mathbf{x}))$, for M. As argued before, $P'(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) = 1[f'_i(\mathbf{pa}_i, \mathbf{u}_i) = v_i]$, where $1[A]$ is the indicator function for an event A. For any not intervened variable, $V_i \notin \mathbf{X}$, the functions f_i and f'_i are the same hence $P'(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) = P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i)$. Since the variables in $\mathbf{X}$ are fixed to constant values $\mathbf{x}$ in M', for every $V_i \in \mathbf{X}$, $\mathbf{Pa}_i = \emptyset$, $\mathbf{U}_i = \emptyset$ and $P'(v_i) = 1[v_i = \mathbf{x}(V_i)]$. This makes intuitive sense because any event where an intervened variable takes a value other than the one fixed by intervention, has probability 0. Putting this observations together leads to

$$P'(\mathbf{v}) = P(\mathbf{v} \mid do(\mathbf{x})) \tag{2.14}$$

$$= \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{X}} 1[\mathbf{v}(V_i) = \mathbf{x}(V_i)] \left(\prod_{V_i \in \mathbf{V} \cap \mathbf{X}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}) \right), \tag{2.15}$$

or, more succinctly to

$$P(\mathbf{v} \mid do(\mathbf{x})) = \begin{cases} \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V} \cap \mathbf{X}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}) & \text{if } \mathbf{v}(\mathbf{X}) = \mathbf{x}, \\ 0 & \text{otherwise} \end{cases}. \tag{2.16}$$

Equation (2.16) is not a function that can be computed from data since $\mathbf{U}$, by definition, is not observed. Nevertheless, it will serve as a logical anchor to connect observational and interventional distributions. Hereon, eq. (2.16) will be referred to as the *latent truncated factorization*.

Chapter 3: A Computational View of Causal Inference

One prominent challenge shared throughout the sciences is to infer cause-and-effect relationships. For instance, to determine how increasing the state's educational budget will bring about changes in the average income of the population, whether exposing subjects to a new advertisement campaign would translate into additional sales revenue, or how patients will react to the decrease of the drug's dosage. Despite the disparate nature of these questions, they evoke the same set of principles and formal machinery, which comes under the rubric of *causal inference* [31, 43].

In general, it is impossible to answer such questions with data alone as further causal knowledge (assumptions) about the underlying model is necessary (but sometimes not sufficient) [39]. Some common assumptions are related to the functions (e.g., linear), their properties (e.g., additive, monotonic), or their functional dependencies.

This chapter introduces a formal notion of *causal inference tasks* and a general framework to solve them. Broadly speaking, a task consists of three main components: (1) a query of interest, (2) a collection of available datasets, and (3) a set of assumptions about the underlying data generating process. Several problems in causal inference can be formulated in those terms.

We will illustrate the proposed framework using three canonical tasks found in the literature. Specifically, we will discuss in this chapter the tasks of: identification from observational data [31, 43] (section 3.3.1), identification from a combination of observational and experimental data [45, 13] (section 3.3.2), and generalization across domains (also known as transportability) [23] (section 3.3.3).

Each inferential task has a specific signature dictated by its components. Based on this formalism, a general solution strategy is constructed incrementally as the tasks are

introduced. Overall, our strategy decomposes the target query and determines whether each of its components can be computed as a function of the given input data, based on the available assumptions. To do this systematically, a new data structure called *causal factor tree* will be defined. This approach provides a unified framework for several problems in nonparametric causal inference and generalizability.

We note that even though the solutions for the tasks discussed in this introductory chapter are already known, this general framework is new and will allow us to solve more challenging inferential tasks in the remainder of the dissertation.

3.1 Causal Inference Tasks

We start by formalizing the signature of a causal inference task consisting of the three components mentioned before.

Definition 7 (Signature, Causal Inference Task)**.** The signature of a causal inference task I is a tuple

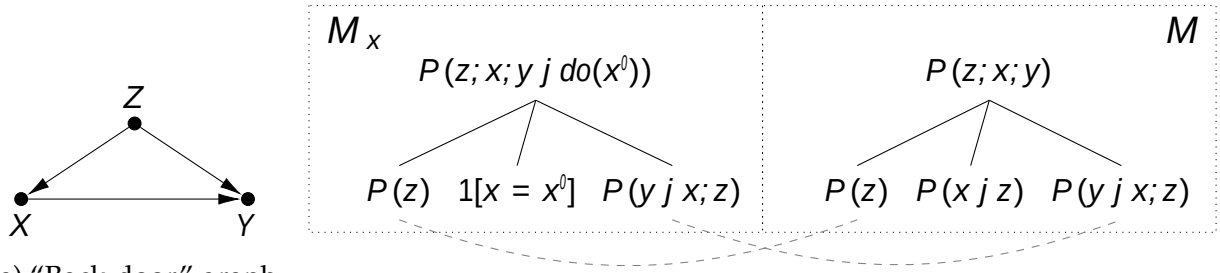
$$I = \langle Q; \mathsf{P}; A \rangle, \tag{3.1}$$

where

- Q is the query — the quantity to be inferred,
- P is the input data — a set of distributions available to the researcher, and
- A is a set of causal assumptions about the underlying SCM.

In this thesis, the causal assumptions will be encoded through the language of causal diagrams. They will facilitate the nonparametric characterization of several tasks. Causal diagrams can be elicited from expert judgment [31] or constructed with the help of structural learning algorithms [46]. The example below illustrates the concept and the role that causal diagrams play in solving causal inference tasks.

Example 2 (A simple causal inference task — Backdoor adjustment)**.** Consider the causal diagram G in figure fig. 3.1(a) and the query $P(y \mid do(x))$, to be identified from $P(\mathbf{V})$.



(a) "Back-door" graph

(b) Conditional probabilities of M_x matched with conditional probabilities of M

Figure 3.1: Conceptual view of the identification of causal effects in the Markovian case. The factors composing the query $P(z; x; y \mid do(x))$ are derived from the factors composing the observational distribution $P(z; x; y)$.

Formally, the signature of this task is

$$I = \langle P(y \mid do(x)); P(\mathbf{V}); G \rangle. \tag{3.2}$$

To solve this instance, first consider the decomposition of the query following the latent truncated factorization (eq. (2.16)):

$$P(z; x; y \mid do(x)) = \sum_{\mathbf{u}} P(z \mid \mathbf{u}_z) P(y \mid x; z; \mathbf{u}_y) P(\mathbf{u}). \tag{3.3}$$

Note that the right-hand side of eq. (3.3) cannot be evaluated immediately since it relies on variables in $\mathbf{U}$, which are not measured by definition (we also do not know their cardinality or form). Still, due to the lack of bidirected edges, it must be the case that $\mathbf{U}_z$; $\mathbf{U}_x$, and $\mathbf{U}_y$ are disjoint and jointly independent. Then, as $\mathbf{U}_x$ does not appear in any factor, it can be summed out. Meanwhile, the sum over $\mathbf{U}_z$ and $\mathbf{U}_y$ can be distributed as

$$P(z; y \mid do(x)) = \left(\sum_{\mathbf{u}_z} P(z \mid \mathbf{u}_z) P(\mathbf{u}_z) \right) \left(\sum_{\mathbf{u}_y} P(y \mid x; z; \mathbf{u}_y) P(\mathbf{u}_y) \right). \tag{3.4}$$

Note each $\mathbf{U}_i$ is independent of $\mathbf{Pa}_i$ for $V_i = Z; Y$, then

$$\sum_{\mathbf{u}_i} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}_i \mid \mathbf{pa}_i) = \sum_{\mathbf{u}_i} P(v_i, \mathbf{u}_i \mid \mathbf{pa}_i) \tag{3.5}$$

$$= P(v_i \mid \mathbf{pa}_i); \tag{3.6}$$

and so

$$P(y \mid do(x)) = \sum_{z,x} P(y \mid x, z) P(z). \tag{3.7}$$

Even though this is a well-known expression, called the backdoor adjustment [47], it is instructive to go through its derivation for understanding the idea behind the proposed framework.

Furthermore, from the latent factorization (eq. (2.12)) of $P(Y, X, Z)$ and the independence between each $\mathbf{U}_i$ and $\mathbf{Pa}_i$, the observational distribution factorizes as:

$$P(y, x, z) = P(y \mid x, z) P(x \mid z) P(z). \tag{3.8}$$

Naturally, each factor on the right-hand side is computable from $P(y, x, z)$ as

$$P(y \mid x, z) = P(y, x, z) = \sum_{y} P(y, x, z); \tag{3.9}$$

$$P(x \mid z) = \sum_{y} P(y, x, z) = \sum_{y,x} P(y, x, z); \text{ and} \tag{3.10}$$

$$P(z) = \sum_{y,x} P(y, x, z) = \sum_{y,x,z} P(y, x, z). \tag{3.11}$$

In summary, eq. (3.7) shows the query as a function of two factors: $P(y \mid x, z)$ and $P(z)$. In turn, eqs. (3.9) to (3.11) re-express those factors as functions of the observational distribution $P(\mathbf{V})$, given as input. Figure 3.1(b) illustrates the overall strategy – to the left we have the query and the factors required to compute it; the right side contains the factors

computable from the data. If we can match every non-trivial factor (e.g., not $1[x = x']$) to the left with one factor to the right, then there exists a mapping from the input to the query. The causal diagram fig. 3.1(a) acts as a catalyst by licensing the decomposition of both the observational and interventional distributions and making the matching approach viable.

The strategy presented in example 2 is straightforward and always successful for tasks where the causal diagram G is Markovian (i.e., has no bidirected arrows). Unfortunately, models representing real-world problems often involve unobserved variables affecting more than one observable variable, a phenomenon that is also known as *latent confounding*. In particular, the factorization of the query and input distributions are not always successful and as simple as in example 2. To witness, consider the following example.

Example 3 (Non-markovian models)**.** The presence of the edge $X \leftrightarrow Y$ in the causal diagram G in fig. 3.2(a) indicates $\mathbf{U}_x$ and $\mathbf{U}_y$ are correlated, or $\mathbf{U}_x \not\perp\!\!\!\perp \mathbf{U}_y$. For observational distributions induced by models described by G, a factorization such as $P(\mathbf{v}) = P(x)P(z \mid x)P(y \mid z)$ does not hold because conditioning on Z does not make Y independent of X; they remain correlated due to the unobserved confounder(s). In particular, the factorization provided by the latent factorization (eq. (2.12)) instantiates for this model as

$$P(\mathbf{v}) = \sum_{\mathbf{u}} P(x \mid \mathbf{u}_x)P(z \mid x, \mathbf{u}_z)P(y \mid z, \mathbf{u}_y)P(\mathbf{u}), \tag{3.12}$$

where $\mathbf{U}_x$ and $\mathbf{U}_y$ are not disjoint or they are correlated.

There is a fundamental difference between eq. (3.12) and a factorization such as eq. (3.8). The former includes unobserved variables and individual factors cannot be computed from data, in contrast to the situation in example 2.

Tian and Pearl [48] generalized the idea used in example 2 to models with latent confounding, expressing the target query in terms of factors computable from the available data. To this end, they introduced the concepts of *c-factor* and *c-component*, leading to a

(a) "Front-door graph" G (b) $G_{\underline{\mathbf{V}}}$

Figure 3.2: The front-door graph [31, p. 81] and the subgraph where each connected component is a c-component.

crisp and convenient decomposition of the distributions induced by models in this class.

Definition 8 (C-factor)**.** For any $\mathbf{C} \subseteq \mathbf{V}$, $Q[\mathbf{C}](\mathbf{v})$ is called the *c-factor* of $\mathbf{C}$ and denotes the function

$$Q[\mathbf{C}](pa(\mathbf{C})) = \sum_{\mathbf{u}(\mathbf{C})} \prod_{V_i \in \mathbf{C}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}(\mathbf{C})), \tag{3.13}$$

where $\mathbf{U}(\mathbf{C}) = \bigcup_{V_i \in \mathbf{C}} \mathbf{U}_i$ and $pa(\mathbf{C})$ is an assignment of the variables $Pa(\mathbf{C})$.[1]

When $\mathbf{C} = \mathbf{V}$, $Q[\mathbf{V}](\mathbf{v})$ matches the right-hand-side of latent factorization equation (eq. (2.12)), hence it is equal to $Q[\mathbf{V}](\mathbf{v}) = P(\mathbf{v})$. For simplicity $Q[\mathbf{C}](pa(\mathbf{c}))$ will be denoted just as $Q[\mathbf{C}]$. For example, the following are some c-factors according to fig. 3.2(a):

$$Q[Y] = \sum_{\mathbf{u}_y} P(y \mid z, \mathbf{u}_y) P(\mathbf{u}_y), \tag{3.14}$$

$$Q[X, Y] = \sum_{\mathbf{u}_y, \mathbf{u}_x} P(x \mid \mathbf{u}_x) P(y \mid z, \mathbf{u}_y) P(\mathbf{u}_y, \mathbf{u}_x), \text{ and} \tag{3.15}$$

$$Q[Z] = \sum_{\mathbf{u}_z} P(z \mid x, \mathbf{u}_z) P(\mathbf{u}_z). \tag{3.16}$$

A c-factor $Q[\mathbf{C}]$ sometimes decomposes into smaller c-factors according to the latent structure of the corresponding subgraph over $\mathbf{C}$, $G[\mathbf{C}]$. Specifically, the variables in a graph (or subgraph) G can be partitioned into sets called *c-components*.

[1] Recall that $Pa(\mathbf{C}) \neq \mathbf{Pa}_{\mathbf{c}}$. The former includes $\mathbf{C}$ itself while the latter refers only to the parents.

Definition 9 (C-component)**.** Two variables $V_i; V_j \in \mathbf{V}$ belong to the same *confounded-component* (for short *c-component*) in a causal diagram G if there exists a path between V_i and V_j made entirely of bidirected arrows.

Equivalently, the c-components of a graph G are the connected components of the subgraph $G_{\underline{\mathbf{V}}}$, where all directed arrows have been removed. For instance, the causal graph in fig. 3.2(a) has two c-components: $\{X; Y\}$ and $\{Z\}$. These are precisely the connected components of the subgraph in fig. 3.2(b). For the causal diagram in fig. 3.1(a), every variable belongs to its own c-component, i.e., $\{Z\}$, $\{X\}$, and $\{Y\}$. The interesting aspect of c-components is that a c-factor associated with a causal diagram G factorizes into smaller c-factors corresponding to the c-components of G. For fig. 3.2(a), this implies that:

$$Q[X; Z; Y] = Q[X; Y]Q[Z]; \tag{3.17}$$

where the first factor in the right-hand side corresponds to the c-component $\{X; Y\}$ and the second to the c-component $\{Z\}$. This equality follows from the observation that the distribution over $\mathbf{U}$ factorizes as $P(\mathbf{u}) = P(\mathbf{u}_x; \mathbf{u}_y)P(\mathbf{u}_z)$. Then $Q[\mathbf{V}]$, given by

$$\sum_{\mathbf{u}} P(x \mid \mathbf{u}_x)P(z \mid x; \mathbf{u}_z)P(y \mid x; z; \mathbf{u}_y)P(\mathbf{u}); \tag{3.18}$$

factorizes as

$$\underbrace{\left(\sum_{\mathbf{u}(X;Y)} P(x \mid \mathbf{u}_x)P(y \mid x; z; \mathbf{u}_y)P(\mathbf{u}(X; Y))\right)}_{Q[X;Y]} \underbrace{\left(\sum_{\mathbf{u}(Z)} P(z \mid x; \mathbf{u}_z)P(\mathbf{u}(Z))\right)}_{Q[Z]}; \tag{3.19}$$

where $\mathbf{U}(X; Y) = \mathbf{U}_X \cup \mathbf{U}_Y$ and $\mathbf{U}(Z) = \mathbf{U}_Z$. The possibility of decomposing c-factors such in eq. (3.19) will play an instrumental role in solving causal inference tasks.

3.2 Causal Factor Trees

In this section, we introduce a data structure called *causal factor tree* that encodes relationships between c-factors. Specifically, it represents whether one c-factor could be *computed* as a function of other c-factors. Later in section 3.3, this will be instrumental for the generalization of the strategy in example 2 to general latent confounding settings.

Definition 10 (Causal Factor Tree (cftree, for short))**.** A rooted polytree[2] T is called a causal factor tree if

(i) every node, including the root, represents a c-factor $Q[\mathbf{C}]$, and

(ii) every node $Q[\mathbf{C}]$ is computable as a function of the nodes at the tail of the edges into it. In other words, there exists some function f such that $Q[\mathbf{C}] = f(fQ[\mathbf{W}] \mid (Q[\mathbf{W}] \not\to Q[\mathbf{C}]) \in T)$.

We will introduce two types of cftrees, one representing the query Q and the other for each input dataset $P \in \mathsf{P}$, which we will call, respectfully, *q-trees* and *d-trees*. A q-tree has arrows into the root, which means that the root is computable as a function of a subset of the leaves. In contrast, a d-tree has arrows going out from the root, which means that any node $Q[\mathbf{C}]$ in the d-tree must be computable via a composition of the functions along the path from the root to $Q[\mathbf{C}]$.

Cftrees will play a central role in our framework since they will serve as computational devices to evaluate whether a query Q is computable from the set of input distributions P. Figure 3.3 illustrates this process, where the q-tree is shown on the top and the d-trees on the bottom. If sufficient factors $Q[\mathbf{D}_1]$, $Q[\mathbf{D}_2]$, $\dots$, $Q[\mathbf{D}_k]$ in the q-tree are found in any of the d-trees, a connection (witnessed by the highlighted paths in the figure) is established between P and the Q, where the latter is computable from the former.

To implement this strategy, the ability to *expand* nodes in a cftree will be required. This is precisely the role of the causal diagram (and the causal assumptions it encodes), to license

[2] A directed acyclic graph whose underline structure is a tree.

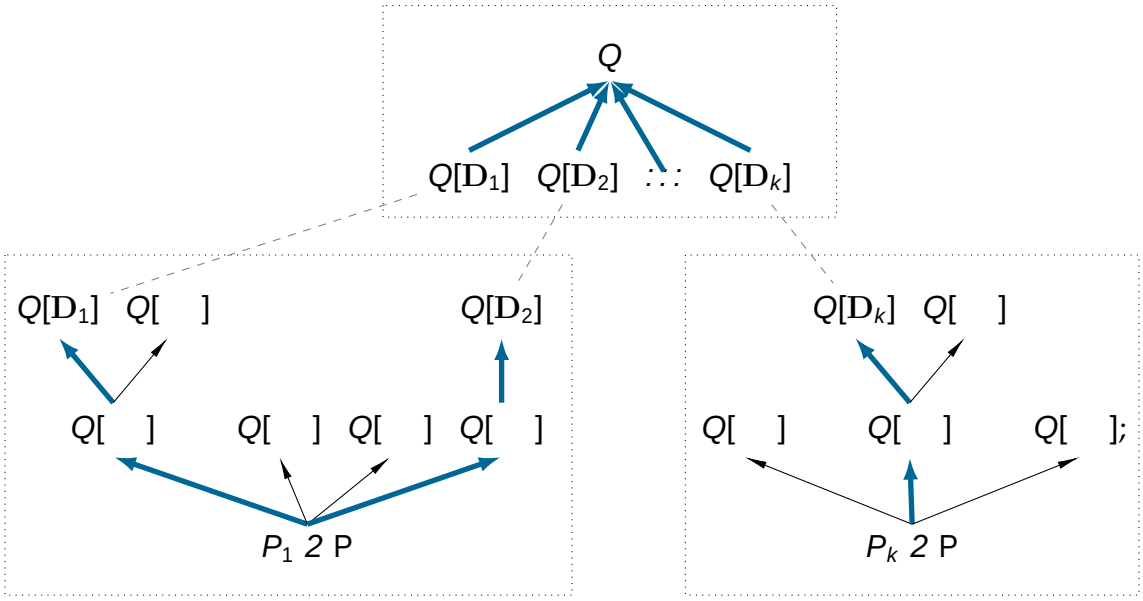


Figure 3.3: General view of the solution strategy for a causal inference task. The query is represented with a q-tree whereas the input distributions are represented with d-trees. If sufficient nodes in the query's tree appear in the input trees, the paths (highlighted) describe a way to compute the query as a function of the inputs.

the validity of the *operators* that will be defined for expanding cftrees. Each operator will be denoted with a greek letter so that the edges in the cftrees can be annotated with them. First, we introduce two basic operators based on two lemmas from [48]: *marginalization* and *independence* among c-factors, as defined next.

Lemma 1 (and operators)**.** *Let endogenous sets* $\mathbf{T}$*;* $\mathbf{C}$ *be such that* $\mathbf{C}$ $\mathbf{T}$ *and let* $\mathbf{C}^0 = \mathbf{T} \cap \mathbf{C}$*. Then, for any cftree* T *with a node* $Q[\mathbf{T}]$*:*

-operator (marginalization): If there is no directed arrow with tail in $\mathbf{C}^0$ *and head in* $\mathbf{C}$ *in the diagram* $G[\mathbf{T}]$*,* $Q[\mathbf{T}]$! $Q[\mathbf{C}]$ *is a valid edge for* T *and it is associated with the function*

$$Q[\mathbf{C}] = \sum_{\mathbf{c}^0} Q[\mathbf{T}]: \tag{3.20}$$

-operator (independence): If there is no bidirected arrow connecting $\mathbf{C}$ *and* $\mathbf{C}^0$ *in the diagram* $G[\mathbf{T}]$*,* $Q[\mathbf{T}]$! $Q[\mathbf{C}]$ *is a valid edge for* T*. The associated function is defined for any*

topological order $T_1 < T_2 <$ $< T_k$ *on* $G[\mathbf{T}]$ *as*

$$Q[\mathbf{C}] = \prod_{T_i \in \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T}]}; \tag{3.21}$$

Furthermore, let $\mathbf{C}_1; \ldots; \mathbf{C}_k$ *be the c-components of* $G[\mathbf{T}]$*, then the set of edges* $\{Q[\mathbf{T}] \; Q[\mathbf{C}_i]\}_{i=1,\ldots,k}$ *are valid for* T *and correspond to the mapping*

$$Q[\mathbf{T}] = \prod_{i=1}^{k} Q[\mathbf{C}_i]; \tag{3.22}$$

Parting from some c-factor $Q[\mathbf{T}]$, the $\sum$-operator produces a c-factor $Q[\mathbf{C}]$ by summing out variables in $\mathbf{T}$ that have no children in $\mathbf{C}$. On the other hand, the $\prod$-operator produces c-factors $Q[\mathbf{C}]$ such that there are no bidirected arrows between $\mathbf{C}$ and other elements in $\mathbf{T}$. Those operators will be used and exemplified in the next section to construct d-trees for input distributions and subsequently q-trees for target distributions.

3.2.1 Generating Cftrees for an Input Distribution

The purpose of d-trees is to determine what c-factors can be computed from the input distributions in a given task. In a cftree, an arrow from one node to another represents that one c-factor is computable from the other. Then, a path from the root of a d-tree to any node implies that such a node can be computed as a function of the root (the input distribution).

The following example discusses d-trees generated for a given observational distribution $P(\mathbf{V})$.

Example 4 (Inducing d-trees for the back-door and front-door graphs)**.** Consider the generation of d-trees for the distribution $P(\mathbf{V}) = P(X; Z; Y)$ in the context of the back-door (fig. 3.1(a)) and the front-door (fig. 3.2(a)) graphs. Such d-trees are rooted at $Q[\mathbf{V}]$ that is equal to $P(\mathbf{V})$.

Figure 3.4: D-trees for $P(X; Z; Y)$ with respect to the back-door (a) and front-door (b) graphs.

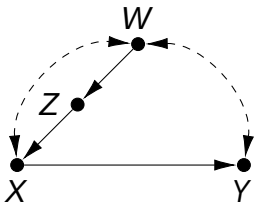
Since the back-door graph has no bidirected edges, three new nodes are produced by the -operator from the root of the d-tree, as shown in fig. 3.4(a). Then, each leaf node is a c-factor with a single variable in it.

For the front-door graph, we could first use the -operator to expand from the root to the nodes $Q[X; Y]$ and $Q[Z]$, given that there are no bidirected arrows between those sets. Then, the -operator licenses the edges $Q[X; Y] \; ! \; Q[X]$ and $Q[X; Y] \; ! \; Q[Y]$, since there are no directed arrows across X and Y in $G[X; Y]$. The resulting d-tree is shown in fig. 3.4(b).

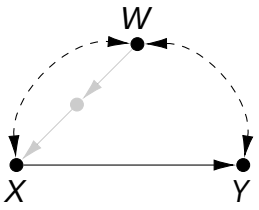
Let us study one more example that includes writing c-factors at the leaves as functions of the root node.

Example 5 (Inducing a d-tree and writing out the c-factors at the leaves)**.** Suppose the distribution $P(\mathbf{V})$ and the "napkin" causal diagram G in fig. 3.5(a) are given as input. The d-tree in fig. 3.5(d) is rooted at $P(\mathbf{V}) = Q[\mathbf{V}]$ and its expansion is licensed by G. The root has two children obtained with -operators. Notice there are no bidirected arrows across $\{W; X; Y\}$ and $\{Z\}$, hence the condition for the -operator is satisfied.

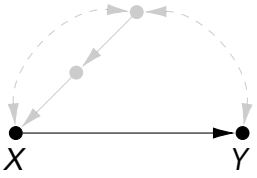
Then, via the -operator, the c-factors $Q[X; Y]$ and $Q[W]$ are computable from the c-factor $Q[W; X; Y]$ because in $G[\{W; X; Y\}]$ (fig. 3.5(b)) there is no arrow from W (the variable being removed) to $\{X; Y\}$. However, $Q[Y]$, for instance, is not computable from $Q[X; Y]$ using the -operator because of the edge $X \; ! \; Y$ in $G[\{X; Y\}]$.



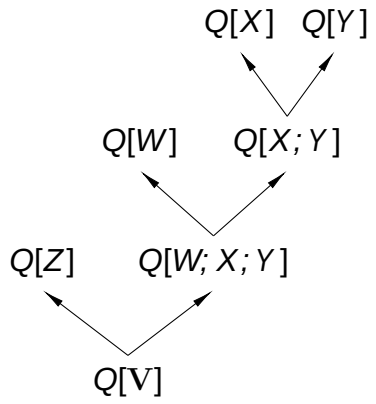
(a) G



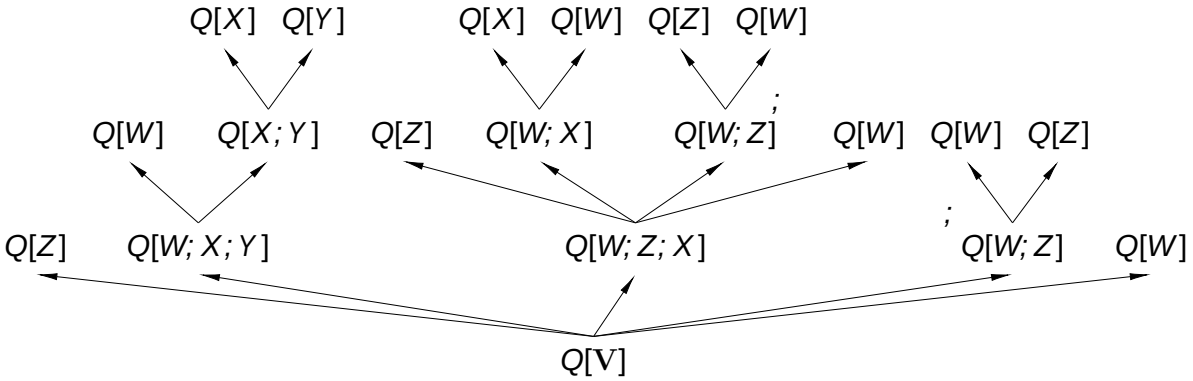
(b) Subgraph $G[W_1; X; Y]$



(c) Subgraph $G[X; Y]$



(d) A d-tree for $P(\mathbf{V})$ induced by the napkin graph



(e) Another d-tree for $P(\mathbf{V})$ and the napkin graph

Figure 3.5: The "napkin" graph, subgraphs associated with the c-factors $Q[W_1; X; Y]$ and $Q[X; Y]$, and two d-trees induced by the observational distribution $P(\mathbf{V})$ and G.

Instead, because there are no bidirected arrows between X and Y in the subgraph $G[\{X, Y\}]$ (fig. 3.5(c)), $Q[Y]$ is computable through the -operator.

Note that as a cftree expands, each directed path entails a series of function involving the node at the tail of each edge and resulting in the node at the head of the edge. For instance, consider the path $Q[\mathbf{V}] \to Q[W, X, Y] \to Q[X, Y]$ in the d-tree of fig. 3.5(d). By the function associated with the -operator, it follows

$$Q[W, X, Y] = \frac{\sum_{z,x,y} Q[\mathbf{V}]}{\sum_{w,z,x,y} Q[\mathbf{V}]} \frac{\sum_{y} Q[\mathbf{V}]}{\sum_{x,y} Q[\mathbf{V}]} \frac{Q[\mathbf{V}]}{\sum_{y} Q[\mathbf{V}]} \tag{3.23}$$

$$= \frac{\sum_{z,x,y} P(\mathbf{v})}{\sum_{w,z,x,y} P(\mathbf{v})} \frac{\sum_{y} P(\mathbf{v})}{\sum_{x,y} P(\mathbf{v})} \frac{P(\mathbf{v})}{\sum_{y} P(\mathbf{v})} \tag{3.24}$$

$$= P(w)P(x \mid z, w)P(y \mid x, z, w). \tag{3.25}$$

Then, by the -operator:

$$Q[X, Y] = \sum_{w} Q[W, X, Y] \tag{3.26}$$

$$= \sum_{w} P(w)P(x \mid z, w)P(y \mid x, z, w). \tag{3.27}$$

Similarly, every node in a d-tree can be computed via the composition of the functions associated with the operators along a path from the root to the node. While expanding a tree, at each node, there could be more than one operator available or different ways to use them. For instance, searching exhaustively through all valid applications of the operators for $P(\mathbf{V})$ and G results in the d-tree shown in fig. 3.5(e).

As the example shows, cftrees are not unique and they could be expanded in different ways. Moreover, depending on the causal diagram, the size of a cftree could be exponential in the number of variables in the diagram. This is why it is important to guide the expansion of the tree based on the factors of interest. As target factors depend on the query,

the following section discusses q-trees generated for queries.

3.2.2 Generating a Q-tree for a Query

While *d-trees* are used for input distributions, *q-trees* are employed to reason about the c-factors required to compute the query of the task. Most of the tasks we will study here have a causal effect as a query, then we will first discuss the translation of interventional distributions into c-factors and their corresponding q-trees.

Consider a model $M_{\mathbf{x}}$ inducing a (mutilated) causal diagram $G_{\overline{\mathbf{X}}}$, and interventional distribution $P(\mathbf{V} \mid do(\mathbf{x}))$. Let $Q[\ \mid do(\mathbf{x})]$ represent the c-factors defined in the context of $M_{\mathbf{x}}$. Then from the latent truncated factorization (eq. (2.16)),

$$P(\mathbf{v} \mid do(\mathbf{x})) = Q[\mathbf{V} \mid do(\mathbf{x})] \tag{3.28}$$

This equality follows from the definition of c-factor in $M_{\mathbf{x}}$. Also by definition, for $X \in \mathbf{X}$,

$$Q[X \mid do(\mathbf{x})](x) = \sum_{\mathbf{u}(X)} P(x \mid do(\mathbf{x}), \mathbf{pa}_X, \mathbf{u}(X))P(\mathbf{u}(X)). \tag{3.29}$$

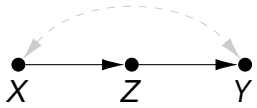
However, $do(\mathbf{x})$ induces a model where f_X is a constant, hence $\mathbf{pa}_X$ and $\mathbf{u}(X)$ are empty in eq. (3.29) and

$$Q[X \mid do(\mathbf{x})](x) = P(x \mid do(\mathbf{x})). \tag{3.30}$$

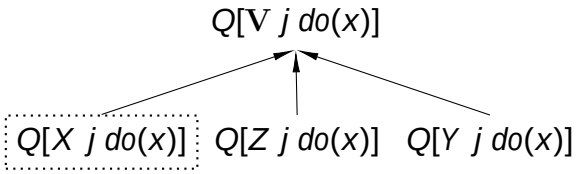
Since the value of X is fixed to $\mathbf{x}(X)$ by the intervention, $P(x \mid do(\mathbf{x}))$ is equal to 1 if x is consistent with $\mathbf{x}(X)$ and 0 otherwise, that is,

$$Q[X \mid do(\mathbf{x})](x) = 1[x = \mathbf{x}(X)].^{3} \tag{3.31}$$

[3] This can be seen as a consequence of a property called *effectiveness* [31, p. 229]. In other words, under an



(a) Interventional causal diagram $G_{\overline{X}}$ for fig. 3.2(a)



(b) A q-tree T induced by $P(\mathbf{v} \mid do(\mathbf{x}))$ and $G_{\overline{X}}$.

Figure 3.6: Example of a q-tree induced by an interventional distribution.

Let us see an example of the generation of q-tree for a query next.

Example 6 (Inducing a q-tree for an interventional query)**.** Consider the front-door graph from fig. 3.2(a) under an intervention $do(X = x)$, as shown in fig. 3.6(a). First, we initialize a q-tree T with root $Q[\mathbf{V} \mid do(x)] = Q[X; Z; Y \mid do(x)]$. Then, using the -operator the root is expanded to three c-factors: $Q[X \mid do(x)]$, $Q[Z \mid do(x)]$, and $Q[Y \mid do(x)]$. This operator is licensed by the lack of bidirected arrows across $X; Z$, and Y in the interventional diagram $G_{\overline{X}}$ (fig. 3.6(a)). While $Q[X \mid do(x)](x')$ is trivially equal to $1[x = x']$, the remaining two factors need to be obtained from the input data in order to compute the query (root of the q-tree) (as discussed in section 3.3).

When the query consists of a marginal causal effect, that is, $Q = P(\mathbf{y} \mid do(\mathbf{x}))$, $\mathbf{Y}$ $\mathbf{V}$, we will first map it to a single c-factor to use as the root of the q-tree. In general, any query $P(\mathbf{y} \mid do(\mathbf{x}))$ can be rewritten in terms of c-factors as follows

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{v} \cap \mathbf{y}} P(\mathbf{v} \mid do(\mathbf{x})) \qquad \text{(Sum over } \mathbf{V} \cap \mathbf{Y}) \qquad (3.32)$$

$$= \sum_{\mathbf{v} \cap \mathbf{y}} Q[\mathbf{V} \mid do(\mathbf{x})] \qquad \text{(c-factor definition)} \qquad (3.33)$$

$$= \sum_{\mathbf{d} \cap \mathbf{y}} Q[\mathbf{D} \mid do(\mathbf{x})]; \qquad (\text{ -operator}) \qquad (3.34)$$

where $\mathbf{D} = An(\mathbf{Y})_{G_{\overline{\mathbf{X}}}}$. The -operator licenses eq. (3.34) because no variable $A \in \mathbf{V} \cap \mathbf{D}$ has arrows pointing to $\mathbf{D}$ in $G_{\overline{\mathbf{X}}}$; otherwise, A would also be an ancestor of $\mathbf{Y}$.

intervention $do(\mathbf{x})$, the probability of observing $X \in \mathbf{X}$ taking a value other than $\mathbf{x}(X)$ is zero.

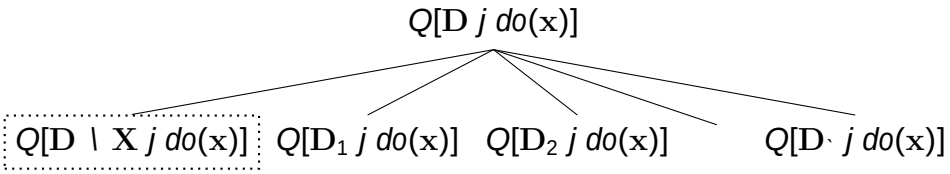


Figure 3.7: Q-tree induced by a marginal causal effect query.

(a) Causal diagram in fig. 3.5(a) under $do(x)$.

(b) Q-tree induced by the distribution $P(y \mid do(x))$ and $G_{\overline{X}}$.

Figure 3.8: Example of a q-tree induced by an interventional distribution.

Based on eq. (3.34), the q-tree is generated with $Q[\mathbf{D} \mid do(\mathbf{x})]$ at the root, as illustrated in fig. 3.7. Then, using the -operator we can expand one node for each c-component of the diagram $G_{\overline{\mathbf{X}}}[\mathbf{D}]$ since there are no bidirected edges across them. For concreteness, let us use this result in the context of an example.

Example 7 (Inducing a q-tree for a marginal interventional query)**.** Consider once again the napkin graph in fig. 3.5(a), this time under the intervention $do(X = x)$. The corresponding causal diagram $G_{\overline{X}}$ is shown in fig. 3.8(a). Also, assume the query of interest is $Q = P(y \mid do(x))$. Following eq. (3.34), we have

$$P(y \mid do(x)) = \sum_{x'} Q[X; Y \mid do(x)]; \tag{3.35}$$

since $An(Y)_{G_{\overline{X}}} = \{X; Y\}$.

Then, we construct a q-tree $T_{P(y|do(x))}$ rooted at $Q[X; Y \mid do(x)]$ (fig. 3.8(b)). Two new nodes are added to the q-tree by applying the -operator, namely, $Q[X; Y \mid do(x)]$ $Q[X \mid do(x)]$ and $Q[X; Y \mid do(x)]$ $Q[Y \mid do(x)]$.

3.3 Solving Causal Identification Tasks with Causal Factor Trees

While every c-factor in a d-tree constructed for an input distribution P is equal to some function of P, in a q-tree for a query Q, the query must be equal to a function of a collection of nodes in the tree. This suggests that solving a causal inference task can be related to matching nodes in those two types of cftrees. The first step in this direction is to establish an explicit connection between interventional and not interventional c-factors.

Consider an intervention $do(\mathbf{x})$ and let $\mathbf{C} \subseteq \mathbf{V}$ be such that $\mathbf{C} \setminus \mathbf{X} = \emptyset$. The c-factor associated with $\mathbf{C}$ in an interventional model is, by definition, given by

$$Q[\mathbf{C} \mid do(\mathbf{x})](pa(\mathbf{C})) = \sum_{\mathbf{u}(\mathbf{C})} \prod_{V_i \in \mathbf{C}} P(v_i \mid do(\mathbf{x}), \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}(\mathbf{C})). \tag{3.36}$$

For every V_i that is not intervened, $P(v_i \mid do(\mathbf{x}), \mathbf{pa}_i, \mathbf{u}_i)$, which only depends on f_i, must be the same as $P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i)$. Therefore,

$$Q[\mathbf{C} \mid do(\mathbf{x})](pa(\mathbf{C})) = \sum_{\mathbf{u}(\mathbf{C})} \prod_{V_i \in \mathbf{C}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) P(\mathbf{u}(\mathbf{C})) \tag{3.37}$$

$$= Q[\mathbf{C}]. \tag{3.38}$$

In other words, the c-factor relative to $\mathbf{C}$ is invariant across pre- and post-interventional distributions. To integrate this type of invariance into our framework, we define a new cftree operator capturing the invariance behind eq. (3.38).

Lemma 2 (operator)**.** *Let $\mathbf{T} \subseteq \mathbf{V}$ be an endogenous set of variables. Then, for any cftree T with a node $Q[\mathbf{T} \mid do(\mathbf{x})]$:*

-operator (regime invariance): If $\mathbf{T} \setminus \mathbf{X} = \emptyset$, $Q[\mathbf{T} \mid do(\mathbf{x})]$ $Q[\mathbf{T}]$ and $Q[\mathbf{T} \mid do(\mathbf{x})]$! $Q[\mathbf{T}]$ are valid edges for T and correspond to the function

$$Q[\mathbf{T}] = Q[\mathbf{T} \mid do(\mathbf{x})]. \tag{3.39}$$

Building on these results, we discuss three canonical problems found in the causal inference literature, formulate them as causal inference tasks (definition 7), and present a general and unifying approach to solving all of them. Specifically, we will consider the following tasks:

1. Identification from observational data (*obs-ID*) (section 3.3.1).
2. Identification from an arbitrary combination of observational and experimental distributions (*g-ID*) (section 3.3.2).
3. Generalization across changing conditions, also called transportability (*g-TR*) (section 3.3.3).

Our framework executes four steps to solve a causal inference task with signature $I = \langle Q; \mathbf{P}; \mathcal{G} \rangle$. First, it generates a q-tree T_Q for the query. Second, it generates a d-tree T_P for each distribution $P \in \mathbf{P}$. Both the construction of T_Q and each T_P are licensed by the assumptions encoded in one or more causal diagrams in $\mathcal{G}$. The third step consists of mapping nodes in T_Q to nodes in some T_P. Finally, if sufficient c-factors from T_Q have been mapped, it constructs an expression for Q as a function of $\mathbf{P}$, as dictated by the paths in the corresponding cftrees. The task fails if one or more required c-factors in T_Q cannot be mapped from the input. This strategy is named C-INFER and is formally described in algorithm 1.

C-INFER is based on four subroutines: GENQUERYTREE, GENINPUTTREE, MAPFACTORS, and COMPOSEQUERY, listed in algorithms 1 to 5. The first step (line 1) in C-INFER is to generate a q-tree T_Q using GENQUERYTREE (algorithm 2) according to the discussion in section 3.2.2 and illustrated in fig. 3.7. Then, for each one of the distributions available as input, a d-tree T_P is generated using GENINPUTTREE (line 3), following the discussion in section 3.2.1. Basically, GENINPUTTREE uses the available cftree operators to "expand" a d-tree rooted at each P. A particular implementation will be discussed below, while solving the first task.

Algorithm 1 C-INFER($I = \langle Q; \mathsf{P}; G \rangle$)

Input: A causal inference task I consisting of a query Q, a set of input distributions P, and a set of causal diagrams G over observable variables $\mathbf{V}$.
Output: Q as a function of P or FAIL.

1: Let T_Q GENQUERYTREE(I). . Generates a q-tree
2: **for each** $P \in \mathsf{P}$ **do**
3: Let T_P GENINPUTTREE($P; I; T_Q$). . Generate d-trees for each input distribution
4: **end for**
5: MAPFACTORS($T_Q; \{T_P\}_{P \in \mathsf{P}}$). . Match c-factors in T_Q to c-factors in some T_P
6: **if** there is a leaf node $Q[\mathbf{D}_j]$ in T_Q such that $\phi(Q[\mathbf{D}_j]) = \emptyset$ **then**
7: **return** FAIL. . Some required c-factor was not found in any d-tree
8: **else**
9: **return** COMPOSEQUERY($T_Q; \emptyset$). . Construct an expression for Q as a function of P
10: **end if**

Algorithm 2 GENQUERYTREE($I = \langle Q; \mathsf{P}; G \rangle$)

Input: A causal inference task I such that $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ and some $G \in \mathcal{G}$ describes the (pre-interventional) SCM associated with Q.
Output: T_Q a q-tree for Q.

1: Let $\mathbf{D} \leftarrow An(\mathbf{Y})_{G_{\overline{\mathbf{X}}}}$.
2: Initialize T_Q with root $Q[\mathbf{D} \mid do(\mathbf{x})]$.
3: Expand T_Q with $Q[\mathbf{D} \mid do(\mathbf{x})] \rightarrow Q[\mathbf{D}_i \mid do(\mathbf{x})]$ for each c-component $\mathbf{D}_i$ of $G_{\overline{\mathbf{X}}}$.
4: **return** T_Q.

Once the query and input cftrees are generated, MAPFACTORS (line 5) simply connects leaf nodes in T_Q with leaf nodes in some T_P. Factors of the form $Q[\mathbf{X} \mid do(\mathbf{x}^{\emptyset})]$ are directly assigned $\mathbb{1}[\mathbf{x} = \mathbf{x}^{\emptyset}]$. If any leaf node of T_Q is not mapped, the procedure fails (line 7). Otherwise, the subroutine COMPOSEQUERY (line 9) returns Q as a function of P using eq. (3.34), and a composition of the functions entailed by the paths between the matched leaves and their root nodes.

In the following, this strategy is instantiated for each of the aforementioned tasks.

3.3.1 Task 1: Identification of Causal Effects from an Observational Distribution

Identifying causal effects from observational data, denoted *obs-ID*, is one of the most well-studied problems in the field and one of the main motivations for the initial development of causal inference methods [31]. In the language developed here, the signature of

Algorithm 3 MAPFACTORS(T_Q; $\{T_P\}_{P \in \mathcal{P}}$)

Input: a q-tree for a query Q and one or more d-trees for each input distribution.
Output: $\Phi : Q[\mathbf{D}_i] \to T$, a mapping from a node $Q[\mathbf{D}_i] \in T_Q$ to a d-tree T such that $Q[\mathbf{D}_i] \in T$.

1: **for each** $Q[\mathbf{Z}_j do(\mathbf{x}')] \in T_Q, \mathbf{Z} \in \mathbf{X}$ **do**
2: Assign $\Phi(Q[\mathbf{Z}_j do(\mathbf{x}')]) = \Phi(1[\mathbf{z}(\mathbf{X}) = \mathbf{x}'])$ $\to Q[\mathbf{Z}_j do(\mathbf{x}')]$. . Trivially assign c-factors of intervened variables
3: **end for**
4: **for each** $Q[\mathbf{D}_j \mid do(\mathbf{x})] \in T_Q, \mathbf{D}_j \setminus \mathbf{X} = \emptyset$ **do**
5: **if** $Q[\mathbf{D}_j \mid do(\mathbf{x})] \in T_P$ for some $P \in \mathcal{P}$ **then**
6: Assign $\Phi(Q[\mathbf{D}_j \mid do(\mathbf{x})]) = T_P$.
7: **end if**
8: **end for**
9: **return** Φ.

Algorithm 4 COMPOSEQUERY(T_Q; Φ)

Input: a q-tree for a query $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ and a mapping of nodes in T_Q to nodes in some T.
Output: A expression for Q.

1: **for each** leaf $Q[\mathbf{C}_i] \in T_Q$ **do**
2: Compute $Q[\mathbf{C}_i]$ as a function of the root of $\Phi(Q[\mathbf{C}_i])$ given by the composition of the functions along the path from the root to $Q[\mathbf{C}_i]$ in that d-tree.
3: **end for**
4: Compute the root of T_Q, $Q[\mathbf{D} \mid do(\mathbf{x})]$, from the leafs.
5: **return** $\sum_{\mathbf{d}\setminus\mathbf{y}} Q[\mathbf{D} \mid do(\mathbf{x})]$.

this task is

$$I_{obs\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})); \{P(\mathbf{V})\}; \{G\}\rangle. \tag{3.40}$$

In other words, the goal is to identify a causal effect from a single observational distribution and the corresponding causal diagram.

Figure 3.9 illustrates the strategy described in algorithm 1 with respect to the classical *obs-ID* task. Specifically, there is one q-tree $T_{P(\mathbf{y}|do(\mathbf{x}))}$ for the query and one d-tree $T_{P(\mathbf{V})}$ for the observational distribution. The task succeeds if every c-factor $Q[\mathbf{D}_i \mid do(\mathbf{x})]$ in $T_{P(\mathbf{y}|do(\mathbf{x}))}$ can be derived in $T_{P(\mathbf{V})}$, and fails otherwise. For concreteness, consider the following examples.

Example 8 (The front-door graph)**.** Let $I = \langle P(y \mid do(x)); P(x, z, y); G\rangle$ where G is the front-

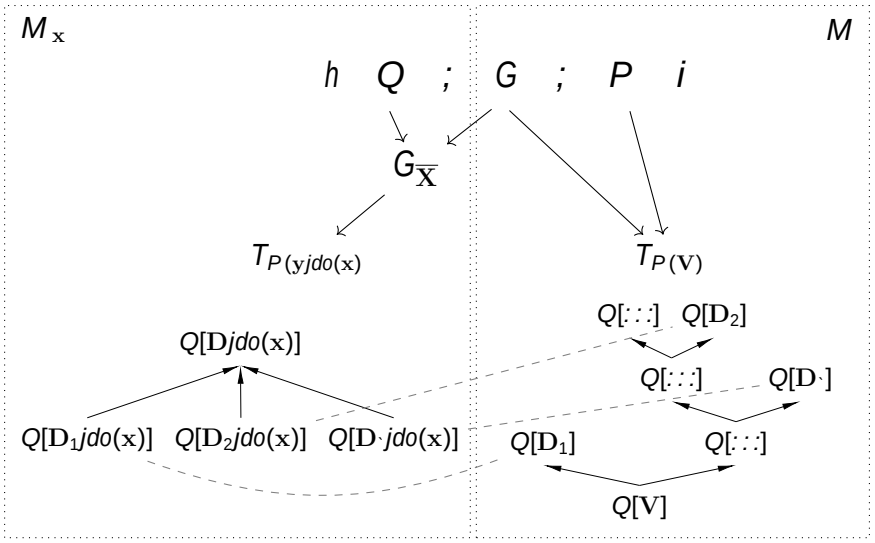


Figure 3.9: Mapping of the c-factors associated with the target query and the c-factors computable from the input distribution.

door graph in fig. 3.2(a). First, GENQUERYTREE generates $T_{P(y|do(x))}$ as shown in fig. 3.10. The q-tree is rooted at $Q[X; Z; Y \mid do(x)]$ since $\{X; Z; Y\}$ is the set $An(Y)_{G_{\overline{X}}}$. Moreover, the root expands to $Q[X \mid do(x)]$ (trivially equal to 1 for x consistent with the intervention), $Q[Z \mid do(x)]$ and $Q[Y \mid do(x)]$.

The next step is to generate a d-tree T_P for the only available distribution $P(\mathbf{V}) = Q[X; Z; Y]$. This is the same d-tree derived in example 4 plus the edges $Q[Z] \rightarrow Q[Z \mid do(x)]$ and $Q[Y] \rightarrow Q[Y \mid do(x)]$ included by GENINPUTTREE (described below in algorithm 5) with the aim of connecting T_Q to T_P. This T_P is shown in the lower section of fig. 3.10. The mapping is successful because all c-factors needed in $T_{P(y|do(x))}$ appear in $T_{P(\mathbf{V})}$.

Finally, the query can be composed according to the paths $Q[X; Z; Y] \rightarrow Q[Z]$ and

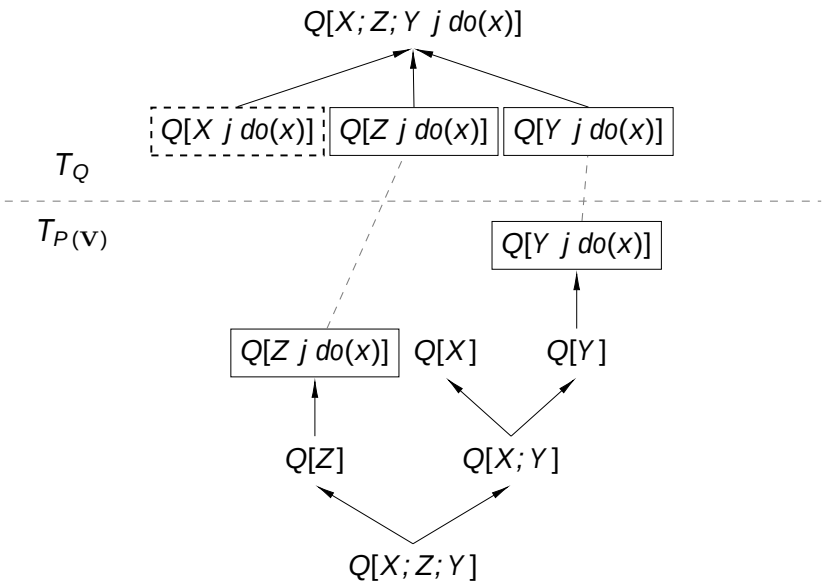


Figure 3.10: Mapping between the query and input cftrees constructed for solving the obs-ID task in the context of the front-door graph.

$Q[X;Z;Y] \;\; ! \;\; Q[X;Y] \;\; ! \;\; Q[Y]$, specifically,

$$Q[Z \mid do(x)] = Q[Z] = \frac{\sum_y Q[X;Z;Y]}{\sum_{z,y} Q[X;Z;Y]} \stackrel{def:}{=} P(z \mid x); \tag{3.41}$$

$$Q[Y \mid do(x)] = Q[Y] = \sum_{x''} Q[X;Y] = \sum_{x''} \frac{\sum_{z,y} Q[X;Z;Y]}{\sum_{x,z,y} Q[X;Z;Y]} \frac{\sum_y Q[X;Z;Y]}{\sum_y Q[X;Z;Y]}$$
$$\stackrel{def:}{=} \sum_{x''} P(x'')P(y \mid x'', z). \tag{3.42}$$

Putting it all together, the target query is equal to

$$P(y \mid do(x)) = \sum_{x',z} Q[X;Z;Y \mid do(x)] \tag{3.43}$$

$$= \sum_{x',z} 1[x' = x](P(z \mid x')) \left(\sum_{x''} P(x'')P(y \mid x'', z) \right) \tag{3.44}$$

$$= \sum_{z} (P(z \mid x)) \left(\sum_{x''} P(x'')P(y \mid x'', z) \right); \tag{3.45}$$

Algorithm 5 GENINPUTTREE(P; $I = \langle Q; \mathbf{V}; G^{(i)}; T_Q \rangle$)

Input: A distribution P, a causal inference task I and a q-tree T_Q.
Output: T_P, a d-tree for P.

1: Initialize T with P at the root . P could be interventional or observational.
2: **for** each node $Q[\mathbf{C} \mid do(\mathbf{x})] \in T_Q$, starting from the root,
at every current node $Q[\mathbf{T} \mid do(\mathbf{z})]$ **do**
3: **if** $\mathbf{C} \cap \mathbf{Z} \neq \emptyset$ **then** give up on $Q[\mathbf{C} \mid do(\mathbf{x})]$. . Target variable intervened in tree
4: **if** $\mathbf{T} = \mathbf{C}$ and $\mathbf{x} = \mathbf{z}$ **then** move to next $Q[\mathbf{C} \mid do(\mathbf{x})]$. Search is done
5: **if** $\mathbf{T} = \mathbf{C}$ and $\mathbf{C} \setminus (\mathbf{X} \cup \mathbf{Z}) = \emptyset$ **then** derive $Q[\mathbf{C} \mid do(\mathbf{x})]$ by -operator. . Change models
6: Let $\mathbf{A} \leftarrow An(\mathbf{C})_{G_{\overline{\mathbf{Z}}}[\mathbf{T}]}$.
7: **if** $\mathbf{A} \neq \mathbf{T}$ **then** use -operator to derive and move to $Q[\mathbf{A} \mid do(\mathbf{z})]$. . Can sum-out variables
8: **if** $G_{\overline{\mathbf{Z}}}[\mathbf{T}]$ has more than one c-component **then** use -operator to derive $Q[\mathbf{W} \mid do(\mathbf{z})]$ where $\mathbf{W}$ is the union of the c-components intersecting $\mathbf{C}$. . Can factorize
9: Give up on $Q[\mathbf{C} \mid do(\mathbf{x})]$. . No operator left
10: **end for**

which is a mapping from the input distribution $P(\mathbf{V})$ to the query Q.[4] This identification result is known as "front-door" adjustment [31, p. 81].

When searching a d-tree for a target c-factor $Q[\mathbf{C}]$, a natural question that arises is how to use the available operators to efficiently determine its computability while avoiding backtracking and possibly exponential time. For instance, in the context of example 5, to generate the d-tree shown in fig. 3.5(d) instead of the much larger one in fig. 3.5(e). To this end, a particular implementation of GENINPUTTREE (algorithm 5), which is based on the *Identify* algorithm [10], is used to address those concerns.

We give a brief description of the d-tree generation strategy. The algorithm expands T_P towards the direction of the c-factors needed by Q (according to T_Q). Starting from the root and at every node, GENINPUTTREE applies a valid cftree operator. Line 3 aborts the search for the target c-factor if any of the target variables have been intervened in the distribution corresponding to the tree. Line 4 determines if the current node is the target c-factor, then the expansion is completed. Line 5 changes the model (determined by the

[4]The index x' is used for the sum added when writing the query through eq. (3.34), and x'' is the index of the sum for the expression obtained for $Q[Y \mid do(x)]$ which is different than x

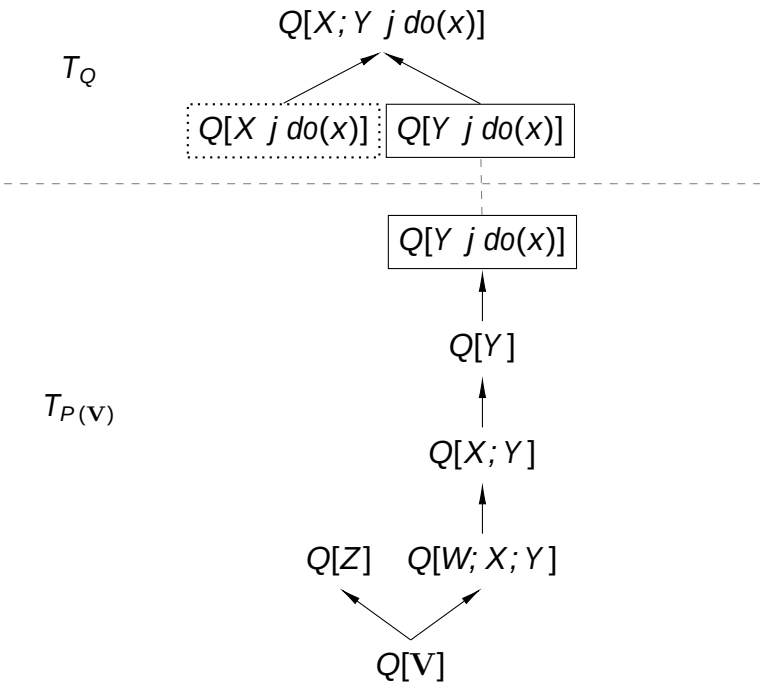


Figure 3.11: cftrees induced for $P(\mathbf{V})$ and the query $P(y \ j \ do(x))$, licensed by the causal diagrams in figs. 3.5(a) and 3.8(a).

intervention) to match the target factor. Line 7 applies if the scope of the c-factor in the current node can be reduced by marginalization. If the current factor contains variables within different c-components than the target variables, line 8 generates a new c-factor where every variable is in the same c-component as the target variables. If none of the previous scenarios applies, the expansion for the current target node is aborted.

Although other strategies are possible, algorithm 5 will be shown to offer the property of completeness and efficiency (theorem 1).

Example 9 (Solving a causal inference task in the napkin graph)**.** Recall the task $I = \langle P(y \ j \ do(x)); P(\mathbf{V}); G \rangle$, where G is the napkin graph in fig. 3.5(a). Following example 7, GENQUERYTREE (algorithm 2) generates $T_{P(yjdo(x))}$, shown in the top part of fig. 3.11. The c-factor $Q[X \ j \ do(x)]$ can be mapped trivially while $Q[Y \ j \ do(x)]$ needs to be matches to some factor in a d-tree.

GENINPUTTREE (algorithm 5) constructs $T_{P(\mathbf{V})}$ aiming for $Q[Y \ j \ do(x)]$ and resulting in the d-tree on the lower half of fig. 3.11. This component is later mapped to its counterpart

in $T_{P(y|do(x))}$ by MAPFACTORS.

Finally, COMPOSEQUERY (algorithm 4) computes $Q[Y \mid do(x)]$ following the path $Q[Y \mid do(x)] \quad Q[Y] \quad Q[X; Y] \quad Q[W; X; Y] \quad Q[\mathbf{V}]$,

$$Q[Y \mid do(x)] = Q[Y] = \frac{Q[X; Y]}{\sum_y Q[X; Y]} = \frac{\sum_w Q[W; X; Y]}{\sum_y \sum_w Q[W; X; Y]} \tag{3.46}$$

$$Q[W; X; Y] = \frac{\sum_{z;x;y} Q[W; Z; X; Y]}{\sum_{w;z;x;y} Q[W; Z; X; Y]} \frac{\sum_y Q[W; Z; X; Y]}{\sum_{x;y} Q[W; Z; X; Y]} \frac{Q[W; Z; X; Y]}{\sum_y Q[W; Z; X; Y]} \tag{3.47}$$

$$\stackrel{def:}{=} P(w)P(x \mid w; z)P(y \mid w; z; x); \tag{3.48}$$

Then, following eq. (3.34), the query distribution can be written as

$$P(y \mid do(x)) = \sum_{x^0} 1[x^0 = x] \frac{\sum_w P(w)P(x \mid w; z)P(y \mid w; z; x)}{\sum_{y;w} P(w)P(x \mid w; z)P(y \mid w; z; x)} \tag{3.49}$$

$$= \frac{\sum_w P(w)P(x \mid w; z)P(y \mid w; z; x)}{\sum_w P(w)P(x \mid w; z)} \tag{3.50}$$

The next example provides an instance where a target c-factor cannot be derived from the d-tree.

Example 10 (A non-identifiable instance)**.** Consider the task with signature $I = \langle P(y \mid do(x)); P(\mathbf{V}); G \rangle$, where G is the causal diagram in fig. 3.12(b). For this task, GENQUERY-TREE constructs the q-tree in fig. 3.12(c), where the c-factor $Q[Y \mid do(x)]$ has to be identified from the input distribution. However, it is impossible to derive this factor from the d-tree associated with $P(\mathbf{V})$ (fig. 3.12(d)). None of the , , or -operators licence the expansion of the root towards another c-factor containing Y. Naturally, GENINPUTTREE will give up (line 9) on $P(\mathbf{V})$ when looking for $Q[Y \mid do(x)]$. Then, C-INFER fails at line 7.

The impossibility of deriving $Q[Y \mid do(x)]$ from $T_{P(\mathbf{V})}$ leads to the question of whether this also implies that the query is not identifiable or the strategy is incomplete. This is indeed the case as we discuss next.

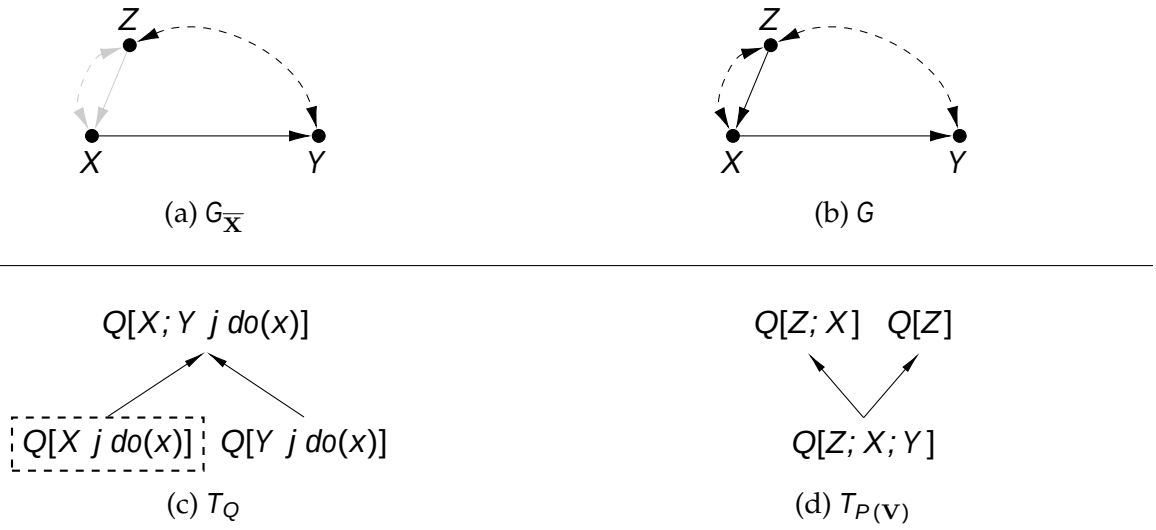


Figure 3.12: Causal diagrams (observational and interventional) and d-trees for a non-identifiable instance.

The failure of C-INFER (algorithm 1) with GENINPUTTREE (algorithm 5) for this instance is not a coincidence. The following result ascertains the necessity (and therefore completeness) of the proposed approach for the identification of causal effects from observational data. In other words, whenever a target factor cannot be derived from the d-tree of the input distribution, the query cannot be identified by any other method (without further assumptions).

Theorem 1 (-, -, and -operators soundness and completeness for obs-ID)**.** *Given a causal inference task $I_{obs\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})); \{P(\mathbf{V})\}; \{G\}\rangle$, the query is identifiable from $P(\mathbf{V})$ and G if and only if* C-INFER *finds a mapping using the , and operators. Moreover, the task is decided in $O(n^2(n+m))$ time following* GENINPUTTREE*, where $n = |\mathbf{V}|$ and m is the number of edges in G.*

When observational data is not sufficient for the identification of a causal effect of interest, there are alternatives based on increasingly rich input datasets. The following section discusses one such scenario where surrogate experiments could help identification.

3.3.2 Identification from a Combination of Experimental and Observational Distributions

Bareinboim and Pearl [13] studied identification from a combination of observational and experimental distributions. This task was called *z-ID* and in contrast to *obs-ID*, it assumes the availability of distributions $P(\mathbf{V} \mid do(\mathbf{Z}^{\emptyset}))$, for every $\mathbf{Z}^{\emptyset} \subseteq \mathbf{Z}$ and some $\mathbf{Z} \subseteq \mathbf{V}$. Later on, Lee, Correa, and Bareinboim [14] studied a relaxation of this setting where surrogate experiments are given for arbitrary sets of intervened variables. This problem is called *g-ID* (general identifiability) and the corresponding task signature can be written as the tuple

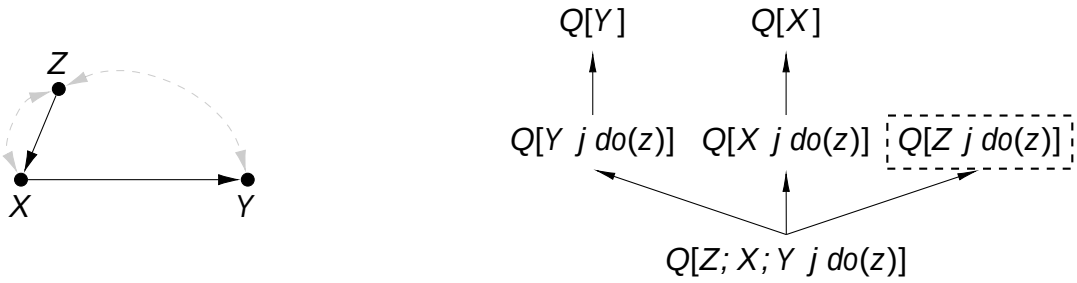
$$I_{g\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})), \mathsf{P}, \mathcal{G}, \mathcal{G}, \mathcal{I} \rangle, \tag{3.51}$$

where $\mathsf{P} = \{P(\mathbf{V} \mid do(\mathbf{z}))\}_{\mathbf{z} \in \mathrm{Val}(\mathbf{Z}), \mathbf{Z} \in \mathcal{Z}}$ and $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_m\}$ with $\mathbf{Z}_i \subseteq \mathbf{V}$. The *z-ID* and *g-ID* tasks differ in two main aspects: experiments for $\mathbf{Z}^{\emptyset} \subseteq \mathbf{Z}_i \in \mathcal{Z}$ are not assumed to be known unless $\mathbf{Z}^{\emptyset}$ itself belongs to $\mathcal{Z}$ (i.e., $P(\mathbf{V} \mid do(\mathbf{z}^{\emptyset}))$ is available), and the observational distribution $\mathbf{Z} = \emptyset$ (no interventions) is not assumed to be available unless $\emptyset \in \mathcal{Z}$.

Example 11 (Identifiability through surrogate experiments)**.** In example 10, the effect $P(y \mid do(x))$ was not identifiable due to the impossibility of deriving the required c-factor $Q[Y \mid do(x)]$ from $P(\mathbf{V})$. If the task is instead $I = \langle P(y \mid do(x)), \{P(\mathbf{V}), P(\mathbf{V} \mid do(\mathbf{Z}))\}, \mathcal{G}, \mathcal{G}_{\mathcal{I}} \rangle$, that is, $P(\mathbf{V} \mid do(z))$ is also given as input, there is an extra source of data and a corresponding d-tree to explore, which could be useful to establish $P(y \mid do(x))$. Figure 3.13(a) and (b) show the causal diagram and d-tree corresponding to the new experimental distribution. The mechanics to compute the tree for $P(\mathbf{V} \mid do(z))$ is exactly the same as for $P(\mathbf{V})$ (as discussed in section 3.2). Note the path $Q[Z, X, Y \mid do(z)] \rightarrow Q[Y \mid do(z)] \rightarrow Q[Y \mid do(x)]$ entails the following mapping:

$$Q[Y \mid do(x)] = Q[Y \mid do(z)] = \frac{Q[Z, X, Y \mid do(z)]}{\sum_{y} Q[Z, X, Y \mid do(z)]} = P(y \mid x, do(z)). \tag{3.52}$$

Hence the query $P(y \mid do(x)) = P(y \mid x, do(z))$ (for any $z \in \mathrm{Val}(Z)$), based on the q-tree in



(a) $G_{\overline{\mathbf{Z}}}$ (b) $T_{P(\mathbf{V}jdo(\mathbf{Z}))}$

Figure 3.13: Causal diagram and d-trees induced by the surrogate experiment.

fig. 3.12(c).

In general, the resulting mapping for Q is not necessarily a function of just one of the input distributions, as illustrated in the next example.

Example 12 (g-Identification with a mix of distributions)**.** Consider the causal inference task with signature $I = \langle P(y \mid do(x)); \mathsf{P} = \{P(\mathbf{V}); P(\mathbf{V} \mid do(Z))\}; \{G\}\rangle$, where G is the diagram in fig. 3.14(b).

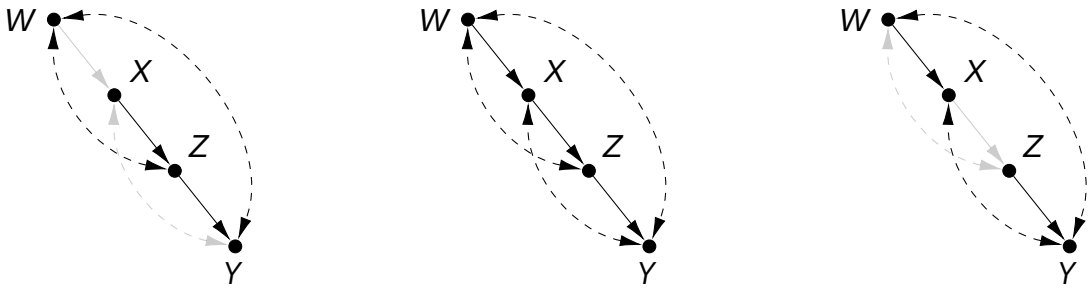
First, GENQUERYTREE produces the q-tree in fig. 3.14(d). The required c-factors for the evaluation of Q are $Q[Z \mid do(x)]$ and $Q[Y \mid do(x)]$. Next, GENINPUTTREE generates the d-trees $T_{P(\mathbf{V})}$ (fig. 3.14(e)) and $T_{P(\mathbf{V}jdo(z))}$ (fig. 3.14(f)). MAPFACTORS finds $Q[Z \mid do(x)]$ in $T_{P(\mathbf{V})}$ and $Q[Y \mid do(x)]$ in $T_{P(\mathbf{V}jdo(\mathbf{Z}))}$, whereas $Q[X \mid do(x)]$ is assigned the usual indicator function. Finally, COMPOSEQUERY maps

$$Q[Z \mid do(x)] = \sum_{w} P(w)P(z \mid x; w); \tag{3.53}$$

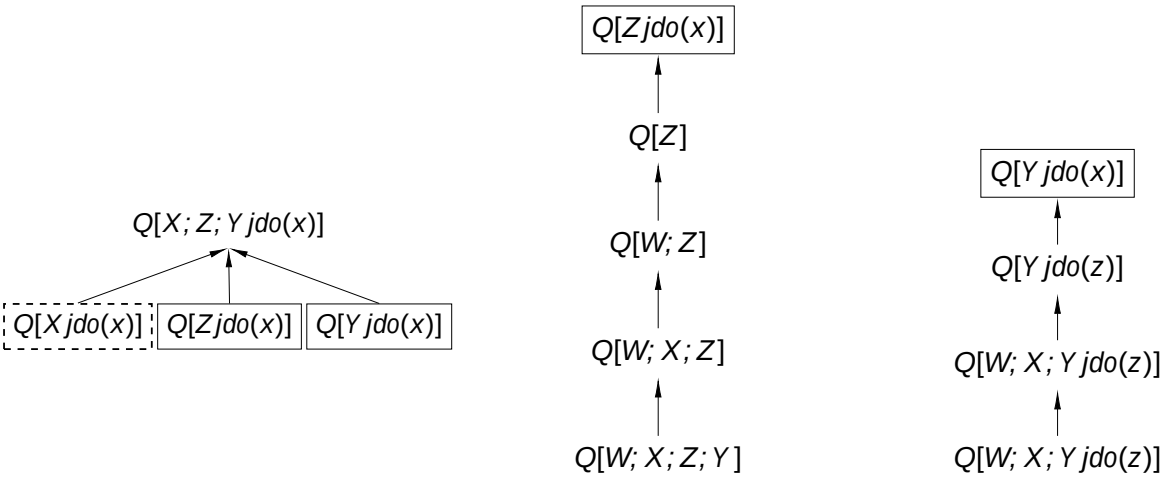
$$Q[Y \mid do(x)] = P(y \mid do(z)); \tag{3.54}$$

and the query can then be written as

$$P(y \mid do(x)) = \sum_{z} \sum_{w} P(w)P(z \mid x; w) \left(P(y \mid do(z))\right). \tag{3.55}$$



(a) $G_{\overline{\mathbf{X}}}$ (b) G (c) $G_{\overline{\mathbf{Z}}}$



(d) $T_{P(y|do(x))}$ (e) $T_{P(\mathbf{V})}$ (f) $T_{P(\mathbf{V}_{jdo(z)})}$

Figure 3.14: Causal diagrams and factors trees used to solve a g-ID instance where the query is identified as a function of more than one input distribution.

Naturally, there are also instances where the query is not identifiable from the given combination of distributions, as shown in the next example.

Example 13 (A non g-identifiable instance)**.** Let $Q = P(y \mid do(x))$ be the query of interest, $\mathsf{P} = \{P(\mathbf{V}); P(\mathbf{V} \mid do(\mathbf{Z}))\}$ and G the causal diagram in fig. 3.15(c). From GENQUERYTREE, we obtain the q-tree in fig. 3.15(b) based on the corresponding causal diagram (fig. 3.15(a)). The c-factors to obtain from the input are $Q[W \mid do(x)]$ and $Q[Z; Y \mid do(x)]$. While the former appears in both $T_{P(\mathbf{V})}$ and $T_{P(\mathbf{V}_{do(z)})}$, the latter is not derivable in any of the d-trees associated with the input distributions (figs. 3.15(d) and 3.15(f)). In turn, this situation implies the non-g-identifiability of this task, as will be stated next.

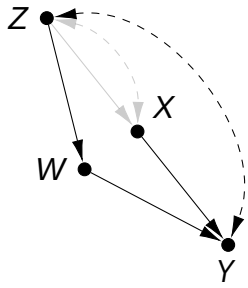
Theorem 2 (-, -, and -operators soundness and completeness for g-ID)**.** *] Given a causal inference task with signature $I_{g\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})); \{P_{\mathbf{z}}(\mathbf{V})\}_{\mathbf{z} \in Val(\mathbf{Z}); \mathbf{Z} \subseteq \mathbf{Z}}; \{G\}\rangle$, for an arbitrary collection of experiments Z, the query is identifiable from P and G if and only if* C-INFER *finds a mapping using the , , and operators. Moreover, the task is decided in $O(n^2(n+m)p)$ time, where $n = |\mathbf{V}|$, m is the number of edges in G and $p = |\mathsf{P}| = |\mathbf{Z}|$.*

When surrogate experiments fail, the next step in the path of identifiability is to borrow data from related (possibly different) domains. The next section covers such a task.

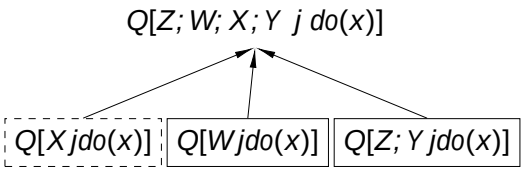
3.3.3 Generalization of Causal Effects Across Environments

Pearl and Bareinboim introduced the problem of transportability [49] which refers to the generalization of causal effects using data from heterogeneous domains. The idea is to use observational and experimental distributions from one or more environments, domains, or populations [23, 25]. Those domains are not the same as the target domain where a query Q is to be estimated.

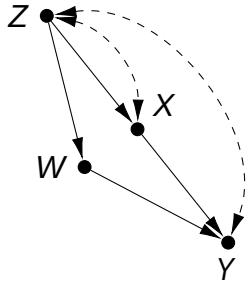
Formally, consider several domains $= \{ ; {}^1; \ldots\}$, each associated to SCMs M , M^1, $\ldots$, where is the target domain.



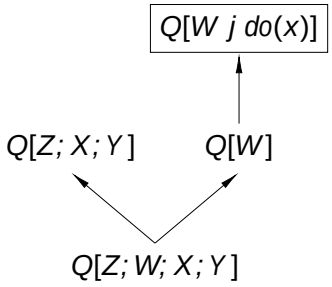
(a) $G_{\overline{\mathbf{X}}}$



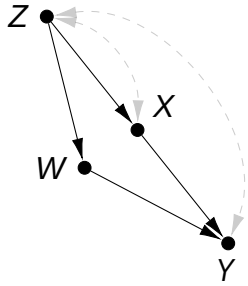
(b) $T_{P(y|do(x))}$



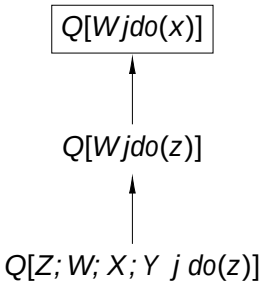
(c) G



(d) $T_{P(\mathbf{V})}$

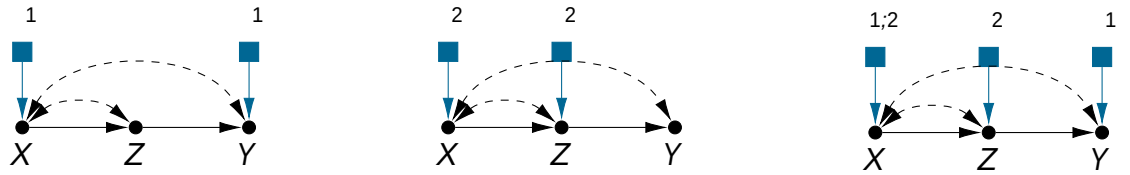


(e) $G_{\overline{\mathbf{Z}}}$



(f) $T_{P(\mathbf{V}|do(z))}$

Figure 3.15: Causal diagrams and d-trees associated with a non-g-ID instance.



(a) G^{Δ_1} (b) G^{Δ_2} (c) G^{Δ}

Figure 3.16: Selection diagrams involving three domains , ^{1} and ^{2}. (a) and (b) show the differences between and ^{1}, ^{2}, respectively. (c) summarizes all differences in a single selection diagram (denoted simply as G^{Δ}).

Lee, Correa, and Bareinboim [50] proposed a general transportability task that generalizes the *g-ID* discussed earlier, called *general transportability* (*g-TR*). This task allows for arbitrary experiments to be given as input in each domain. To formally study this task, first, we need to encode the disparities and commonalities in the causal mechanisms of the different domains involved in the analysis. To precisely specify these assumptions, the following definitions are used.

Definition 11 (Domain Discrepancy)**.** Let ^{a} and ^{b} be domains associated, respectively, with SCMs M^a and M^b, conforming to causal diagrams G^a and G^b. Denote by $^{a;b}$ $\mathbf{V}$ a set of variables such that, for every V_i 2 $^{a;b}$, there is possibly a discrepancy between domains, namely, $f_i^a \not\equiv f_i^b$ or $P^a(\mathbf{U}_i) \not\equiv P^b(\mathbf{U}_i)$.

For simplicity, the set of discrepancies between and ^{i}, $^{;i}$, is written simply written as ^{i} with $=$;. Moreover, the discrepancies are encoded graphically with a specialization of causal diagrams called *selection diagrams*.

Definition 12 (Selection Diagram [23])**.** Given a causal diagram $G^i = \langle \mathbf{V}; \mathbf{E} \rangle i$ and domain discrepancies ^{i}, let $\mathbf{T} = \{T_v \mid \exists_{i=1}^{n} V \; 2 \; ^i\}$ be called *selection variables*. Then, a *selection diagram* G ^{i} is defined as a graph $\langle \mathbf{V} \cup \mathbf{T}; \mathbf{E} \cup \{T_v \to V\}_{T_v \in \mathbf{T}} \rangle i$.

For example, suppose there are three domains ^{1}; ^{2}, and . Figure 3.16 (a), (b) [51, fig. 2] represents two selection diagrams comparing the target domain with ^{1} and ^{2}, respectively. Moreover, G^{Δ} in fig. 3.16(c) is a superimposition of the selection

diagrams summarizing the differences between and each of the source domains where the distributions in P come from.[5]

The signature of the task corresponding to g-TR is given by

$$I_{g\text{-}TR} = P(\mathbf{y} \mid do(\mathbf{x})); \mathsf{P}; \mathcal{G}^{\Delta} ; \tag{3.56}$$

where $\mathsf{P} = \{P^i(\mathbf{V} \mid do(\mathbf{z}_j)) \mid \mathbf{z}_j \in \text{Val}(\mathbf{Z}_j); \mathbf{Z}_j \in Z^i\}_{Z^i \in \mathcal{Z}}$, $\mathcal{Z} = \{Z^i \mid i \in \}$, with each $Z^i = \{\mathbf{Z}_1; \mathbf{Z}_2; \ldots\}, \mathbf{Z}_j \quad \mathbf{V}$, specifying distributions available in domain ^{i}.

In this scenario, the input could be much richer; there is not only data in the target domain , but also from other domains, which could be observational or interventional. The number of d-trees increases accordingly. The query of interest can be written in terms of c-factors in yet the goal is to determine the conditions under which c-factors from other domains can be generalized to . We define a new cftree operator to capture this notion of invariance. Let $Q^a[\mathbf{C}]$ represent a c-factor for the set $\mathbf{C}$ in an SCM M^a. Accordingly, the following operator can be defined.

Lemma 3 (operator)**.** *Let* $\mathbf{T} \quad \mathbf{V}$ *be an endogenous set of variables. Then, for any cftree* T *with a node* $Q^a[\mathbf{T}]$*:*

-operator (domain invariance): If $\mathbf{T} \setminus {}^{a;b} = ;$, $Q^a[\mathbf{T}] \quad Q^b[\mathbf{T}]$ *and* $Q^a[\mathbf{T}] \quad ! \quad Q^b[\mathbf{T}]$ *are valid edges for* T *and corresponds to the mapping*

$$Q^a[\mathbf{T}] = Q^b[\mathbf{T}]: \tag{3.57}$$

The -operator simply captures the notion that whenever the mechanism for a set of variables $\mathbf{C}$ are invariant across domains ^{a} and ^{b}, this implies that the corresponding c-factors will be invariant as well. Let us consider an example to illustrate the use of this operator.

[5] Although $\mathcal{G}^{\Delta}$ is a superimposition of selection diagrams, we will often refer to it just as "the selection diagram", for simplicity.

Algorithm 6 GENINPUTTREE(P; $I = \langle Q; \mathbf{P}; G\ i; T_Q$)

Input: A distribution P, a causal inference task I and a q-tree T_Q.
Output: a d-tree T for P.

1: Initialize T with $P = Q^b[\mathbf{V}\ j\ do(\mathbf{x})]$ at the root.
2: **for** each node $Q^a[\mathbf{C}\ j\ do(\mathbf{x})]\ 2\ T_Q$, starting from the root, at every current node $Q^b[\mathbf{T}\ j\ do(\mathbf{z})]$ **do**
3: **if** $\mathbf{C} \setminus \mathbf{Z} \not\in$; or $\mathbf{C} \setminus$ $^{a;b} \not\in$; **then** give up on $Q[\mathbf{C}\ j\ do(\mathbf{x})]$. . Target variable intervened or different in tree
4: **if** $\mathbf{T} = \mathbf{C}$, $\mathbf{x} = \mathbf{z}$ and $a = b$ **then** move to next $Q^a[\mathbf{C}\ j\ do(\mathbf{x})]$. Search is done
5: **if** $\mathbf{T} = \mathbf{C}$ and $\mathbf{C} \setminus (\mathbf{X} \cap \mathbf{Z}) =$; **then** derive $Q^b[\mathbf{C}\ j\ do(\mathbf{x})]$ by -operator. . Model is different
6: **if** $\mathbf{T} = \mathbf{C}$ and $\mathbf{C} \setminus$ $^{a;b} =$; **then** derive $Q^a[\mathbf{C}\ j\ do(\mathbf{x})]$ by -operator. . Domain is different
7: Let $\mathbf{A}$ $An(\mathbf{C})_{G_{\overline{\mathbf{Z}}}[\mathbf{T}]}$.
8: **if** $\mathbf{A} \not\in \mathbf{T}$ **then** use -operator to derive and move to $Q^b[\mathbf{A}\ j\ do(\mathbf{z})]$. . Can sum-out variables
9: **if** $G_{\overline{\mathbf{Z}}}[\mathbf{T}]$ has more than one c-component **then** use -operator to derive $Q^b[\mathbf{W}\ j\ do(\mathbf{z})]$ where $\mathbf{W}$ is the union of the c-components intersecting $\mathbf{C}$. . Can factorize
10: Give up on $Q^a[\mathbf{C}\ j\ do(\mathbf{x})]$. . No operator left
11: **end for**

Example 14 (Transporting c-factors across domains)**.** Recall the selection diagram G^{Δ} in fig. 3.16(c). The domain discrepancies are $^1 = \{X; Y\}$ and $^2 = \{X; Z\}$. Then, the -operator licenses the edges $Q^1[Z] \ ! \ Q\ [Z]$ and $Q^2[Y] \ ! \ Q\ [Y]$ with corresponding mappings

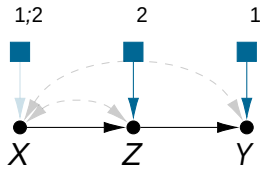
$$Q\ [Z] = Q^1[Z]; \text{and} \tag{3.58}$$

$$Q\ [Y] = Q^2[Y] \tag{3.59}$$

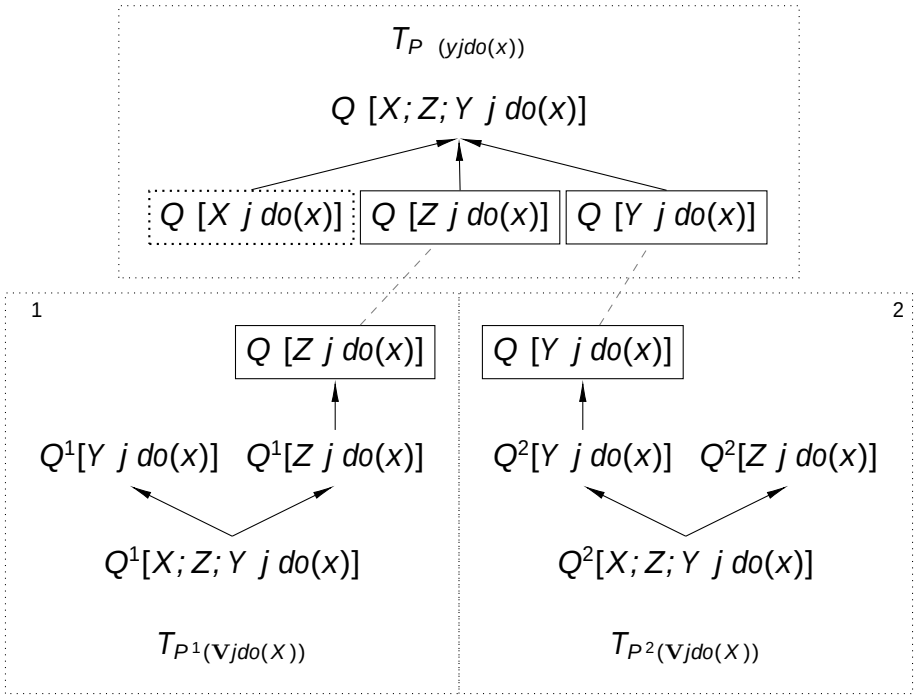
for any cftree associated with G^{Δ}.

With the addition of the -operator, GENINPUTTREE needs to be refined to take into account the domain of the distribution and the new operator (see line 6 in algorithm 6).[6] Moreover, a search on a tree related to domain ^{b} could be aborted if one or more target variables could have differences (belong to $^{a;b}$) concerning a target domain ^{a}.

[6] Changes appear underlined.



(a) $G^{\Delta}_{\overline{X}}$



(b) Cftrees for the query and input distributions within the different domains.

Figure 3.17: Selection diagram under intervention $do(x)$ and an overview of the matching process of the c-factors across distributions and domains via the cftrees.

The following example illustrates the use of all cftree operators defined so far.

Example 15 (A simple transportability task)**.** Consider the causal inference task $I = \langle P\ (y\ |\ do(x);\ \{P^1(\mathbf{V}\ |\ do(x)); P^2(\mathbf{V}\ |\ do(x))\}; G^{\Delta}\rangle$, where G^{Δ} is the selection diagram in fig. 3.16(c).

GENQUERYTREE produces the q-tree $T_{P\ (y|do(x))}$ in fig. 3.17(b) where the c-factors $Q\ [Z\ |\ do(x)]$ and $Q\ [Y\ |\ do(x)]$ need to be identified from available input distributions. GENINPUTTREE generates the d-trees $T_{P^1(\mathbf{v}|do(x))}$ and $T_{P^2(\mathbf{v}|do(x))}$ in the same figure. Each d-tree represents a distribution in a different domain, and the -operator licenses the connection between the c-factors $Q^1[Z\ |\ do(x)]$ and $Q^2[Y\ |\ do(x)]$ with their counterparts

in , as established by MAPFACTORS.

Finally, COMPOSEQUERY returns Q as

$$P(y \mid do(x)) = \sum_{x';z} 1[x = x']P^1(z \mid do(x))P^2(y \mid do(x); z) \quad (3.60)$$

$$= \sum_{z} P^1(z \mid do(x))P^2(y \mid do(x); z): \quad (3.61)$$

Note that the first factor in eq. (3.61) can be computed from the interventional distribution in ^{1} while the second can be computed from the interventional distribution in ^{2}.

Below, there is one more example based on a causal diagram in [25].

Example 16 (A g-TR identification task)**.** Consider a causal inference task $I = \langle P(y \mid do(x)); \mathsf{P}; G^{\Delta}\rangle$, where $\mathsf{P} = \{P(\mathbf{V}); P^1(\mathbf{V} \mid do(Z_2)); P^2(\mathbf{V} \mid do(Z_1))\}$ and G^{Δ} is shown in fig. 3.18(c) (from [25]).

First, we generate T_Q (fig. 3.18(b)) where the c-factors to be obtained from the input are $Q[Z_1]$, $Q[Z_3]$, $Q[R]$ and $Q[W; Y]$. The c-factors $Q[W; Y]$ and $Q[Z_1]$ appear in $T_{P(\mathbf{V})}$ (fig. 3.18(d)), whereas $Q[Z_3]$ and $Q[R]$ do not belong to that d-tree. They may, however, be found in the two d-trees given by the distributions in ^{1} and ^{2} (figs. 3.19(b) and 3.19(d)). For c-factors of to be derived from those d-trees, the -operator has to be used at some point. This means $Q[R]$ cannot be derived from $P^2(\mathbf{V} \mid do(Z_1))$ because $R \in$ ^{2} (shown in G^{Δ} as a square node marked with ^{2} pointing to R), even if $Q^2[R]$ is a node in $T_{P^2(\mathbf{V} \mid do(Z_1))}$. It can, however, be derived in $T_{P^1(\mathbf{V} \mid do(Z_1))}$ as shown in fig. 3.19(b). Similarly, the c-factor $Q[Z_3]$ cannot appear in a tree for ^{1} because $Z_3 \in$ ^{1}. Still, $Q[Z_3]$ can be mapped from $T_{P^2(\mathbf{V} \mid do(Z_1))}$ as in fig. 3.19(d). Finally, COMPOSEQUERY outputs the transported query as

$$P(y \mid do(x)) = \sum_{z_1;z_2;w;r} Q[Z_1]Q[Z_3]Q[R]Q[W; Y]; \quad (3.62)$$

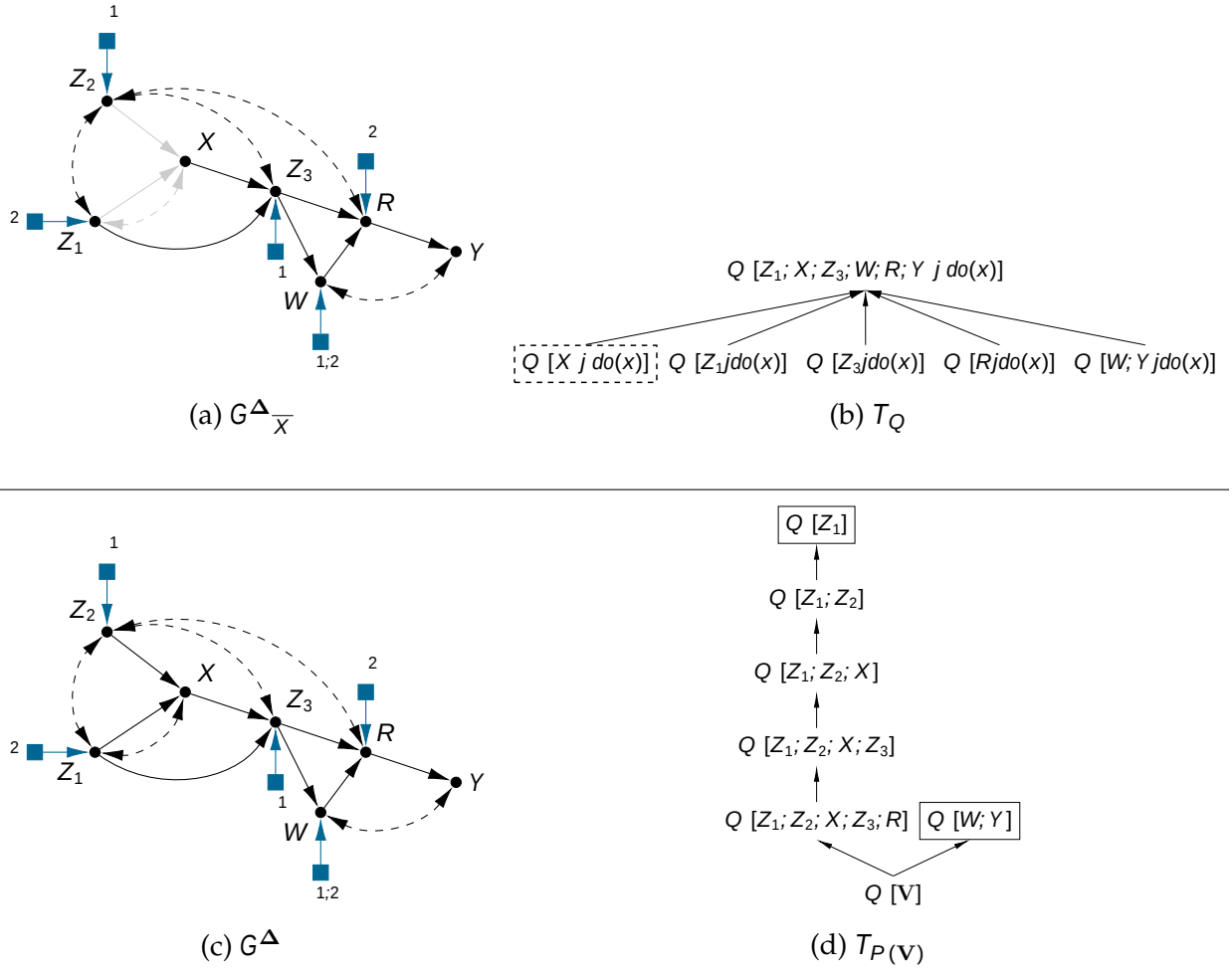
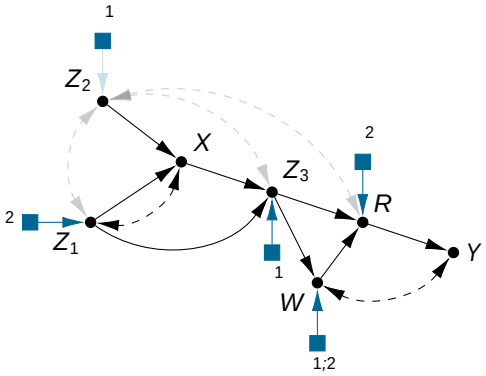
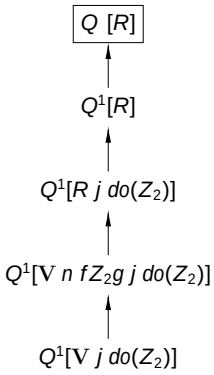


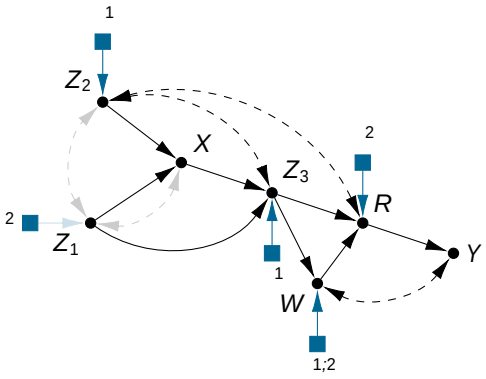
Figure 3.18: Selection diagrams before and after the intervention $do(x)$, together with the q-tree and the d-tree for $P\ (\mathbf{V})$.



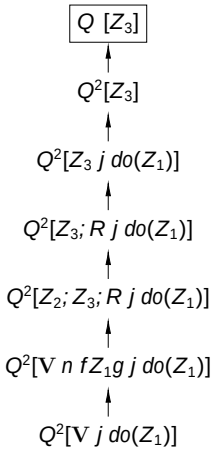
(a) $G^{\Delta}_{\overline{Z_2}}$



(b) $T_{P^1(\mathbf{V}|do(Z_2))}$



(c) $G^{\Delta}_{\overline{Z_1}}$



(d) $T_{P^2(\mathbf{V}|do(Z_1))}$

Figure 3.19: Selection diagrams and d-trees for experimental distributions in two different domains.

where

$$Q\ [Z_1] = P\ (z_1) \tag{3.63}$$

$$Q\ [Z_3] = \sum_{z_2;r} P^2(z_2\ j\ do(z_1))P^2(z_3\ j\ do(z_1); z_2; x)P^2(r\ j\ do(z_1); z_2; x; z_3; w) \tag{3.64}$$

$$Q\ [R] = P^1(r\ j\ do(z_2); z_1; x; z_3; w) \tag{3.65}$$

$$Q\ [W; Y\ ] = P\ (w\ j\ z_1; z_2; x; z_3)P\ (y\ j\ z_1; z_2; x; z_3; w; r): \tag{3.66}$$

With the addition of the -operator to the set of cftree operators, the approach from algorithm 1 becomes complete for the g-TR task.

Theorem 3 (-, -, -, and -operators soundness and completeness for g-TR)**.** *Given a causal inference task with signature $I_{g\text{-}TR}$, the query is transportable from* P *and* G^{Δ} *if and only if* C-INFER *finds a mapping using the , , , and operators. Moreover, the task is decided in* $O(n^2(n + m)p)$ *time, where* $n = |\mathbf{V}|$*,* m *is the number of edges in* G*, and* $p = |\mathbf{P}|$*.*

3.4 Summary

In this section, we described increasingly complex causal inference tasks, each one subsuming the previous one. Specifically, the set of input distributions P first included only observational data, then experimental data, and finally data from several domains. The causal diagram was also refined into a selection diagram (in some sense, overlapping of multiple causal diagrams) to account for the multiple domains. Table 3.1 lists these tasks, their signatures (Q; P and G), the cftree operators used to solve them and their sufficient and necessary characterization (S&N).

The query Q, however, remained the same —a causal effect with a *do* intervention. In the next chapter, we generalize this aspect by considering a more general class of interventions, where the manipulated variables could follow a conditional or stochastic policy, based on other observable variables.

	Name	Q	P	G	Ops	S&N
obs-ID	Identification from observational data	$P(\mathbf{y}\|do(\mathbf{x}))$	$P(\mathbf{V})$	G	′ ′	,
g-ID	General Identifiability	$P(\mathbf{y}\|do(\mathbf{x}))$	$fP(\mathbf{V}\|do(\mathbf{Z}))g_{\mathbf{Z}2Z}$	G	′ ′	,
g-TR	General Transportability	$P^{*}(\mathbf{y}\|do(\mathbf{x}))$	$fP^{i}(\mathbf{V}\|do(\mathbf{Z}_j))g_{\mathbf{Z}_j 2Z^i 2Z}$	G^{Δ}	′ ′ ,	,

Table 3.1: Summary of the causal inference tasks discussed in sections 3.3.1 to 3.3.3.

Chapter 4: Causal Inference with Soft Interventions[1]

The interventions induced by the *do* operator, also called *hard* or *atomic* interventions, are considered primitives in causal analysis. However, interventions may consist of complex policies where a variable X is set to follow a conditional or stochastic relationship depending on other variables in the system. In the following, these interventions will be referred to as *soft interventions*.

For instance, consider the causal diagram in fig. 4.1(a), and let X represent the choice of *smoking* or not, W *age*, Z a set of risk factors leading to *a tendency to smoke* (e.g., peer pressure, education, SES, exposure to advertisement, psychological age), and Y the development or not of *lung cancer*. The interventional distribution $P(Y \mid do(X = x))$ describes the behavior of Y when X is fixed to x (smoking or not) *regardless of* Z or any other confounding factors. This is graphically illustrated in fig. 4.1(b). Furthermore, the difference between two specific levels of the intervention, $P(Y \mid do(X = 1)) \quad P(Y \mid do(X = 0))$, measures the causal changes of Y that are due to the deliberate variations of X.

While the intervention $do(X = 0)$ describes with mathematical precision a counterfactual world where smoking is banned from society, it is unlikely, in practice, that a policy could be implemented such that cigarettes would be completely wiped out from the streets. In other words, we could eventually predict the effect of this new, idealized policy, even though it may be hard to implement in reality. Some authors [52, 53, 54] have raised concerns regarding the *do* operator. They suggest, for instance, that most practical policies are nonatomic, hence a causal analysis based on *do* interventions is too idealistic because it assumes a precise and surgical ability to replace the causal mechanisms of the intervened variables, which is not achievable in practice. On the other hand, Pearl [31, p. 106] has

[1]This chapter is based on the papers [32] and [33].

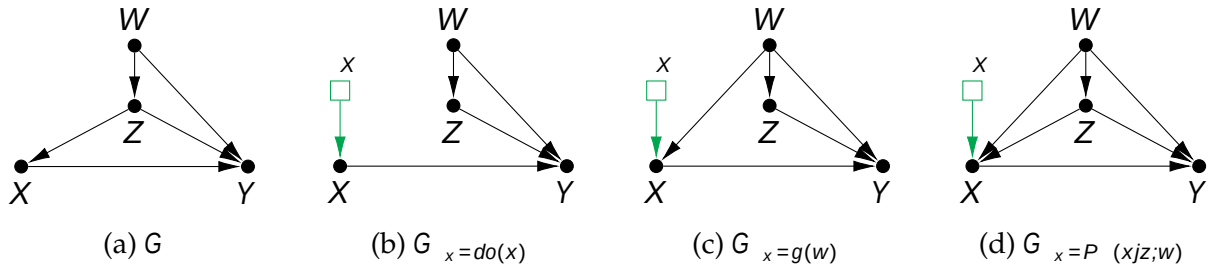


Figure 4.1: (a) original causal diagram G. (b), (c), and (d) show the causal diagrams after an atomic, conditional, and stochastic intervention, respectively.

argued that these concerns emerge from conflating the mathematical definition of causal effect with the feasibility of its implementation and that despite its ideal nature, the *do* operator permits us to analyze more complex interventions.

For instance, policy-makers could contemplate a more strict regulation on underage smoking and higher taxes on cigarettes sales that could be set in place. A sensible question in this context could be "what is the effect of a policy that inhibits smoking in people under 21 years of age, by 90%?" Contrary to an atomic intervention entailing that a 100% decrease in smoking should be enforced for this group, this intervention suggests that the underlying mechanism for X should be replaced with a "softer" mechanism.

Even though deciding the identifiability of complex interventions has been studied and acknowledged in the literature (for a summary see [31, Ch. 4]), there is still work to be done. For instance, [9] studied the effect of interventions in longitudinal settings where the decision in each time step is dependant on the previous ones, which was called *conditional plans*. Further, other works investigated the effect of *stochastic* interventions, where the original causal mechanism of the treatment variable is replaced with a new known function [42, 55, 56, 57].

Despite the high level of sophistication and generality achieved for reasoning with atomic interventions, there are glaring differences with the nonatomic case. For instance, there exist no counterpart for *do*-calculus [31, Sec. 3.4] in the nonatomic case nor general results on identifiability from experimental distributions produced by soft interventions, akin to the results discussed in section 3.3.3. In this chapter, we develop a general, symbolic,

and algorithmic treatment for the identifiability of soft interventions from a combination of observational and experimental distributions, based on the algorithmic framework introduced in chapter 3. More specifically:

- **Symbolic characterization.** Section 4.2 introduces a set of inference rules to reason about the effect of general types of intervention, called -calculus. Furthermore, we provide a syntactical method for deriving and verifying claims about such interventions, given a causal graph.

- **Algorithmic solution.** Section 4.3 generalizes the algorithmic framework introduced in chapter 3 to solve tasks involving nonatomic interventions both in the query and the input's side of the analysis. For instance, by asking for the effect of a conditional or stochastic policy and using data generated by soft interventions.

4.1 Soft Interventions and SCMs

Although several results and discussions exist for the identification of the effect of soft interventions [9, 42, 58, 56, 59, 57], the classical definitions of *submodel* (definition 2) and *potential response* (definition 3) cannot directly represent policy-like interventions within the SCM semantics introduced in chapter 1. To formalize the notion of a model resulting from a soft intervention the following definition is proposed.

Definition 13 (Interventional Model)**.** Let $M = \langle \mathbf{V}, \mathbf{U}, F, P(\mathbf{U}) \rangle$ be a causal model, $\mathbf{X}$ a set of variables in $\mathbf{V}$, and $f_{\mathbf{x}} = \{f_x\}_{X \in \mathbf{X}}$ be a set of functions, where each $f_x : \text{Val}(\mathbf{Z}_x) \to \text{Val}(\mathbf{U}_x) \to \text{Val}(X)$ is function from a set of random variables $\mathbf{U}_x$, with distribution $P(\mathbf{U}_x)$, and set of endogenous variables $\mathbf{Z}_x \subseteq \mathbf{V} \setminus De(X)$. An interventional model $M_{\mathbf{x}}$ derived from M is the causal model

$$M_{\mathbf{x}} = \langle \mathbf{U} \cup \{\mathbf{U}_{\mathbf{x}}\}, \mathbf{V}, F_{\mathbf{x}}, P(\mathbf{U}, \mathbf{U}_{\mathbf{x}}) \rangle, \tag{4.1}$$

where

$$F_{\mathbf{x}} = \{f_i : V_i \in \mathbf{X}\} \cup \{f_{\mathbf{x}}\}; \tag{4.2}$$

$$\mathbf{U}_{\mathbf{x}} = \bigcup_{X \in \mathbf{X}} \mathbf{U}_X; \tag{4.3}$$

$$P(\mathbf{U}; \mathbf{u}_{\mathbf{x}}) = P(\mathbf{U})P(\mathbf{u}_{\mathbf{x}}). \tag{4.4}$$

In contrast to a submodel (definition 2), the intervened variables in interventional models are not generally fixed to constants and vary according to the specified functions $f_{\mathbf{x}}$. The behavior of the variables in $\mathbf{X}$ could be constant, deterministic, or stochastic depending on how the functions make use of other observables and the randomness encoded in $\mathbf{u}_{\mathbf{x}}$.

Hereon, only interventional SCMs without cyclic functional dependencies are considered. In other words, it is assumed that the causal diagram of any interventional model must be acyclic.

The original model M and an interventional model $M_{\mathbf{x}}$ represent two different *regimes*. The first corresponds to the natural generative process, without any modification. The second represents the same system once the mechanisms associated with $\mathbf{X}$ have been intervened [60, 42, 61]. In the following, *regime indicators* as defined in [62, 42] are used to represent different types of interventions. The regime indicator $\mathcal{I}_X = s$ corresponds to a model $M_{\mathcal{I}_X}$, where the function f_X is replaced following a strategy s. Such model has its own (interventional) causal diagram, $G_{\mathcal{I}_X}$, and induces a distribution $P(\mathbf{V}; \mathcal{I}_X)$. For instance, with respect to the causal diagram G in fig. 4.1(a), the causal diagram in fig. 4.1(b) is the usual $G_{\overline{X}}$ corresponding to a $do(x)$ intervention. Moreover, following the convention in [42], $G_{\mathcal{I}_{\mathbf{X}}}$ is annotated with a node $\mathcal{I}_{X_i}$ and an edge $(\mathcal{I}_{X_i} \to X_i)$ for every $X_i \in \mathbf{X}$ to indicate the targets of intervention. Figure 4.1(c) is the diagram for a intervention on X that depends on W (example 17), whereas fig. 4.1(d) is the causal diagram after some

Type	Strategy	Function
Idle	$\tau_X = \emptyset$	$\hat{f}_x = f_x$
Atomic/*do*	$\tau_X = x$	$\hat{f}_x = do(x)$, for some $x \in \text{Val}(X)$
Conditional	$\tau_X = g(\hat{\mathbf{pa}}_x)$	$\hat{f}_x = g(\hat{\mathbf{pa}}_x)$
Stochastic/Random	$\tau_X = \hat{P}(X \mid \hat{\mathbf{pa}}_x)$	$\hat{f}_x$ s.t. $\hat{P}(x \mid \hat{\mathbf{pa}}_x) = \sum_{u_x} P(\hat{f}_x(\hat{\mathbf{pa}}_x, u_x) = x) P(u_x)$

Table 4.1: Intervention strategies considered. Each row contains a type of intervention, its representation using the regime indicator, and the corresponding $\hat{f}_x$.

intervention on X that depends on both W and Z (as discussed in example 18).

In particular, depending on the intervention strategy, the function $\hat{f}_X$ could accept as arguments the values of observable variables other than the original parents $\mathbf{Pa}_x$. Moreover, since the variables in $\mathbf{U}_x$ are not observed, it is also assumed that any new policy does not depend on any unobservable in $\mathbf{U}$. It is assumed then that for any $\hat{X}$, the set $\hat{\mathbf{U}}_x$ is disjoint and independent of the set $\mathbf{U}$ of the original model (i.e., the natural regime). Accordingly, the set of observable and unobservable parents of X in $M_{\hat{X}}$ are denoted $\hat{\mathbf{Pa}}_x$ and $\hat{\mathbf{U}}_x$. Naturally, this also implies G_{τ_X} may not be a subgraph of G, as it always occurs with *do*-interventions. This is indeed the case in fig. 4.1(b) but not in fig. 4.1(c) or fig. 4.1(d).

4.1.1 Representing Different Interventional Strategies

Qualitatively different types of interventions can be modeled by assigning different *strategies* to the regime indicator τ_X using the construct discussed above. We list in table 4.1 the types of interventions that will be considered. Specifically, the *idle* intervention represents the natural state of the system; *atomic* or *do* interventions replace the function f_X with a constant, while *conditional* ones replace it with a deterministic function of some observables $\hat{\mathbf{pa}}_x$. The *stochastic* type sets the new $\hat{f}_x$ such that the variable X will follow a pre-specified distribution $\hat{P}(X \mid \hat{\mathbf{pa}}_x)$. Whenever clear from the context, the strategy assigned to τ_X will be omitted. Also, $P(\mathbf{V}; \tau_X = \emptyset)$ is simply written as $P(\mathbf{V})$. For a set $\mathbf{X} \subseteq \mathbf{V}$, let $\tau_{\mathbf{x}} = \{\tau_{X_1}, \ldots\}$ represent an intervention affecting each f_{X_i} for every $X_i \in \mathbf{X}$.

Example 17 (Conditional Intervention)**.** In the context of a tutoring program, consider again the causal diagram in fig. 4.1(a). Let W represent the previous GPA of a student, Z student's motivation, taking X after-hours tutoring (or not), and Y the GPA at the end of the term. Currently, students seek tutoring voluntarily, which depends on their motivation. Given the limited amount of resources, the school is considering making after-hours tutoring mandatory for students with low GPAs and offering this service only to them. The proposed intervention can be encoded as $_X = g(w)$, where $g(w) = 1$ if W is low GPA, and 0 otherwise. Graphically, this policy is represented by the diagram in fig. 4.1(c), where X now depends on W, not on Z, as it did in the observational regime and dataset.

Example 18 (Stochastic Intervention)**.** Consider a different interpretation of fig. 4.1(a), where X represents choosing to *smoke*, W *age*, Z a set of risk factors leading to *tendency to smoke* (e.g., peer pressure, education, SES, psychological age), and Y the development or not of *lung cancer*. The government is interested in the effect of a new policy that could reduce by 90% smoking on people under 21 years old. This could be represented with a stochastic intervention $_x = P^{\flat}(x \mid w; z)$ such that $P^{\flat}(X = 1 \mid W < 21; z) = (0{:}1) \quad P(X = 1 \mid z)$, for every z, which is show in fig. 4.1(d).

There is an important difference between atomic and conditional/stochastic interventions due to the deterministic nature of the former. The difference has to do with conditioning on the intervened variable, as discussed next.

Remark 1 (Conditioning on the intervened variable)**.** In contrast to more general interventions, $do(\mathbf{x})$ interventions implicitly conditions on $\mathbf{X} = \mathbf{x}$, formally,

$$P(\mathbf{y} \mid do(\mathbf{x})) = P(\mathbf{y};\ _{\mathbf{X}} = do(\mathbf{X} = \mathbf{x})) \tag{4.5}$$

$$= \sum_{\mathbf{x}'} P(\mathbf{y} \mid \mathbf{x}';\ _{\mathbf{X}} = do(\mathbf{X} = \mathbf{x}))P(\mathbf{x}';\ _{\mathbf{X}} = do(\mathbf{X} = \mathbf{x})) \tag{4.6}$$

$$= P(\mathbf{y} \mid \mathbf{x};\ _{\mathbf{X}} = do(\mathbf{X} = \mathbf{x})). \tag{4.7}$$

Equation (4.5) follows by definition of $do(\mathbf{x})$ strategy for $_{\mathbf{X}}$, eq. (4.6) is obtained by conditioning on $\mathbf{X}$, and eq. (4.7) is immediate since under $_X = do(\mathbf{X}=\mathbf{x})$, the probability $P(\mathbf{x}'; \ _{\mathbf{X}} = do(\mathbf{X} = \mathbf{x}))$ is nonzero only for $\mathbf{x}' = \mathbf{x}$.[2] In general, however, the distributions $P(\mathbf{y}; \ _{\mathbf{X}})$ and $P(\mathbf{y} \mid \mathbf{x}; \ _{\mathbf{X}})$ need not match.

Interestingly, whereas atomic interventions always reduce the model structure (i.e., to a subgraph), a policy-maker could envision new conditional/stochastic policies that take into account a wide range of covariates, and that do not match the observational regime or previous policies (as discussed in example 17, 18).

4.1.2 Relationship between Soft and Atomic Interventions

Considering that the identification of the effect of atomic interventions is a well-understood problem, a natural question is if the problem of identifying soft interventions can be solved leveraging the results for atomic interventions. In [31, Sec. 4.2], Pearl proposes a reduction of the effect of interventions $^1_X = g(\mathbf{Z})$ and $^2_X = P(X \mid \mathbf{Z})$ to the effect of $do(x)$ interventions, as follows:

$$P(\mathbf{y}; \ ^1_X) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{z}; \ ^1_X) P(\mathbf{z}; \ ^1_X) \tag{4.8}$$

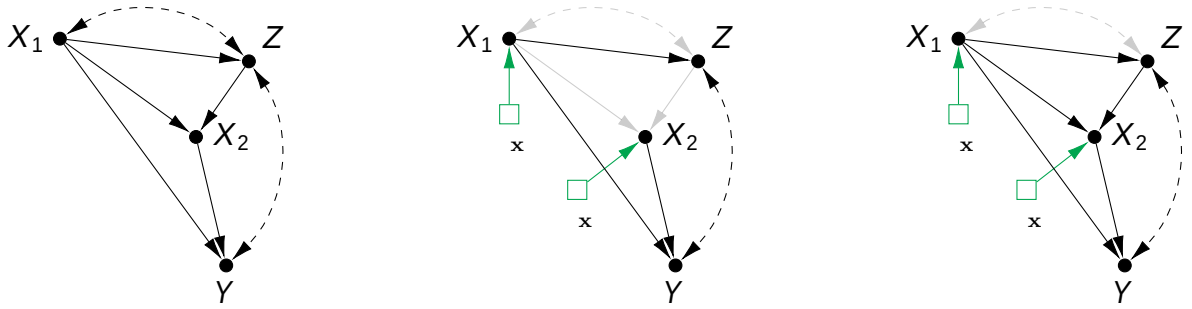
$$= \sum_{\mathbf{z}} P(\mathbf{y} \mid x; \mathbf{z})|_{x=g(\mathbf{z})} P(\mathbf{z}) \text{ and} \tag{4.9}$$

$$P(\mathbf{y}; \ ^2_X) = \sum_{\mathbf{z}; x} P(\mathbf{y} \mid x; \mathbf{z}; \ ^2_X) P(\mathbf{z}; \ ^2_X) \tag{4.10}$$

$$= \sum_{\mathbf{z}; x} P(\mathbf{y} \mid x; \mathbf{z}) P(x \mid \mathbf{z}) P(\mathbf{z}). \tag{4.11}$$

In both cases, the ability to write $P(\mathbf{z}; \ ^1_X)$ and $P(\mathbf{z}; \ ^2_X)$ as $P(\mathbf{z})$ follows from the fact that $\mathbf{Z}$ must contain no descendant of X, hence any intervention on X does not have an effect on the distribution of $P(\mathbf{z})$. Once $\mathbf{Z}$ is fixed, in the case of ^1_X, X follows the mapping $x = g(\mathbf{z})$. For ^2_X, X will take a particular value x according to the distribution $P(x \mid \mathbf{z})$.

[2] This can also be seen through the lens of the property of effectiveness [31, p. 229].



(a) G (b) $G_{\mathbf{x}}$ (regime 1) (c) $G_{\mathbf{x}}$ (regime 2)

Figure 4.2: Causal diagrams associated with example 19.

Although this reduction is intuitive and sound, it lacks a formal justification following from the SCM semantics. In particular, this reduction cannot be derived using the rules do-calculus [31].

Moreover, note we stated these conditions for a single treatment variable X. When we consider joint interventions affecting a set of variables $\mathbf{X}$, the soft intervention on one variable could be a function of a descendant of another intervened variable. Therefore, it is not so clear if the reasoning that led to eq. (4.9) and eq. (4.11) still holds in such a case. The following example from the same reference illustrates how the lack of a formal statement or rules specific to soft interventions could lead to confusion.

Example 19 (Do-calculus can be misleading for soft interventions)**.** Consider the the dynamic plan problem used in [31, Sec. 4.4] associated with the causal diagram in fig. 4.2(a). The goal is to assess the distribution $P(\mathbf{y};\ _{\mathbf{x}})$ in two hypothetical regimes:

Regime 1 the values of $(X_1; X_2)$ have been fixed to $(x_1; x_2)$, in standard do-form, i.e., $_{\mathbf{x}} = f\ _{X_1} = do(x_1);\ _{X_2} = do(x_2)g$. The corresponding $G_{\mathbf{x}}$ is shown in fig. 4.2(b).

Regime 2 the value of X_1 is fixed to x_1 and X_2 is set conditionally on X_1 and Z based on a function $g(x_1; z)$. That is, $_{X_1}$ is as before $_{X_1} = do(x_1)$, and $_{X_2} = g(x_1; z)$. The corresponding $G_{\mathbf{x}}$ is shown in fig. 4.2(c).

The first scenario (regime 1) is identifiable by the rules of do-calculus (see appendix A.2) as

follows:[3]

$$
\begin{aligned}
& P(y|do(x_1), do(x_2)) \\
&= P(y|x_1, do(x_2)) && \text{(Rule 2: } (Y \perp\!\!\!\perp X_1 \mid X_2) \text{ in } G_{\overline{X_2}\underline{X_1}}) && (4.12)\\
&= \sum_z P(y|x_1, do(x_2), z) P(z|x_1, do(x_2)) && \text{(Condition on } Z) && (4.13)\\
&= \sum_z P(y|x_1, do(x_2), z) P(z|x_1) && \text{(Rule 3: } (Z \perp\!\!\!\perp X_2 \mid X_1) \text{ in } G_{\overline{X_2}}) && (4.14)\\
&= \sum_z P(y|x_1, x_2, z) P(z|x_1). && \text{(Rule 2: } (Y \perp\!\!\!\perp X_2 \mid X_1, Z) \text{ in } G_{\underline{X_2}}) && (4.15)
\end{aligned}
$$

In other words, the effect of the atomic policy is identifiable from observational data.

We now consider regime 2 associated with the query $P(y \mid do(x_1), do(X_2 = g(x_1, z)))$. First, we can apply the second rule of do-calculus to exchange $do(X_1 = x_1)$ with $X_1 = x_1$, exactly as in the first step of regime 1, licensed by $(Y \perp\!\!\!\perp X_1 \mid X_2)$ in $G_{\overline{X_2}\underline{X_1}}$.

Nevertheless, the extra edge $Z \to X_2$ in fig. 4.2(c) (regime 2) indicates that X_1 and Y may be dependent conditional on X_2, hence the same derivation strategy would not work. More concerning is the fact that the effect of $\mathbf{x}$ (regime 2) is not identifiable at all from $P(\mathbf{V})$. We show this by constructing two SCMs M_0 and M_1 that induce the same $P(\mathbf{V})$ and G, yet the differ on $P(y; \mathbf{x})$.

Let $M^i = \langle \mathbf{V}, \mathbf{U}, F^i = \{f_{x_1}, f_z^i, f_{x_2}, f_y\}, P(\mathbf{U})\rangle, i = 0, 1$, where $\mathbf{V} = \{X_1, Z, X_2, Y\}$, the

[3] Refer to [31, p. 120] for a more detailed derivation and discussion about the example.

set of unobservables is $\mathbf{U} = \{U_{x_1z}, U_z, U_{x_2}, U_y\}$ with all $\mathbf{U}$ are binary and the functions

$$f_{x_1} : X_1 \quad U_{x_1z}; \tag{4.16}$$

$$f_z^0 : Z \quad \begin{cases} X_1 \quad U_z & \text{if } X_1 = U_{x_1z} \\ 0 & \text{otherwise} \end{cases}; \qquad f_z^1 : Z \quad \begin{cases} X_1 \quad U_z & \text{if } X_1 = U_{x_1z} \\ 1 & \text{otherwise} \end{cases}; \tag{4.17}$$

$$f_{x_2} : X_2 \quad X_1 \quad Z \quad U_{x_2}; \tag{4.18}$$

$$f_y : Y \quad X_1 \quad X_2 \quad U_y; \tag{4.19}$$

where $\quad$ is the *xor* operator. In terms of $P(\mathbf{U})$, U_{x_1z}, U_z, and U_{x_2} are fair coins, whereas $P(U_y = 1) = 3/4$,

Clearly, both models induce the given causal diagram G. Moreover, without any intervention, X_1 is always equal to U_{x_1z} in both models, so f_z^0 and f_z^1 coincide in the observational regimes and the models induce the same $P(\mathbf{V})$.

Let $g(X_1, Z) = X_1 \quad Z$ and let the interventional models be denoted as $M^0{}_{X_1 \; X_2}$ and $M^1{}_{X_1 \; X_2}$. Those models are identical to M^0 and M^1 except for the replacement f_{x_1} and f_{x_2} with

$$\not{p}_{x_1} : X_1 \quad x_1; \tag{4.20}$$

$$\not{p}_{x_2} : X_2 \quad X_1 \quad Z: \tag{4.21}$$

Under intervention $f_{X_1} = do(x_1), \; _{X_2} = g(X_1, Z)g$, these models induce $P^0(Y = 1; \; _{X_1X_2}) = 5/8, P^1(Y = 1; \; _{X_1X_2}) = 3/8$, proving the non-identifiability of the effect. (See appendix E.1 for details on these calculations.)

This example was deemed identifiable in [31, p. 120] since do-calculus was used to infer $P(y|do(x_1), do(X_2 = g(x_1, z))) = P(y|x_1, do(X_2 = g(x_1, z)))$, which is not true in general. Since do-calculus only considers the removal of edges, this is taken as an indication that the rules of *do*-calculus are not suitable to reason about interventions where the graphical

structure of the intervened diagram is contingent on the type of intervention.

Alternatively to the do-calculus strategy, Tian [56] proposed a reduction of the effect of a soft intervention on a singleton or a set of variables to the effect of a do intervention, proved its soundness, and conjectured its completeness.

Next, we state a slightly more general result in terms of identifiability from arbitrary experimental distributions, which is also necessary and implies also the necessity of Tian's reduction, and, therefore, its completeness.

Theorem 4. *Let $\mathbf{Y}; \mathbf{X} \quad \mathbf{V}$ be any two sets of variables, and let $_{\mathbf{X}}$ be an atomic, conditional, or stochastic intervention. Then, the distribution of $\mathbf{Y}$ under $_{\mathbf{X}}$ can be written as*

$$P(\mathbf{y};\ _{\mathbf{X}} = \ _{\mathbf{X}}) = \sum_{\mathbf{d}n\mathbf{y}} P(\mathbf{d}\,n\,\mathbf{x};\ _{\mathbf{X}} = \mathbf{x}) \prod_{X \in \mathbf{X} \setminus \mathbf{D}} P(x \mid \mathbf{pa}_x;\ _{\mathbf{X}} = \ _{\mathbf{X}}). \tag{4.22}$$

Moreover, this effect is identifiable from $hG; Zi$ if and only if $P(\mathbf{d}\,n\,\mathbf{x};\ _{\mathbf{X}} = \mathbf{x})$ is identifiable from $hG; Zi$, where $\mathbf{D} = An(\mathbf{Y})_{G\ _{\mathbf{X}}}$.

Although theorem 4 offers a reduction that can be further solved using do-calculus, this first step remains unaccounted for in the context of this set of inference rules. Moreover, while do-calculus allow us to read various constraints involving interventional distributions from the causal diagram, this reduction cannot be used for such a purpose in the context of soft interventions. The next section introduces a new set of rules that will not only allow us to read constraints involving soft interventions, but also derive the effect of soft interventions using only the rules and probability axioms from beginning to end.

4.2 A Calculus for Soft Interventions[4]

In this section, we introduce *-calculus*, a set of inference rules in the spirit of *do*-calculus [45] capable of handling both atomic and nonatomic interventions. After introducing the

[4]This section is based on the paper [32].

rules, we compare the two calculi (section 4.2.1) and then provide a more elaborate example (section 4.2.2).

Theorem 5. *[Inference Rules — -calculus] Let G be a causal diagram compatible with an SCM M, with endogenous variables* $\mathbf{V}$. *For any disjoint subsets* $\mathbf{X}; \mathbf{Y}; \mathbf{Z} \quad \mathbf{V}$, *two disjoint subsets* $\mathbf{T}; \mathbf{W} \quad \mathbf{V} \; n \; (\mathbf{Z} \; [\; \mathbf{Y})$ *(i.e., possibly including* $\mathbf{X}$*), the following rules are valid for any intervention strategies* $_{\mathbf{X}}$; $_{\mathbf{Z}}$, *and* $^{()}_{\mathbf{Z}}$ *such that* $G_{\mathbf{X}\,\mathbf{Z}}$ *and* $G_{\mathbf{X}\,()_{\mathbf{Z}}}$ *have no cycles:*

Rule 1 (Insertion/Deletion of observations):

$$P(\mathbf{y} \; j \; \mathbf{w}; \mathbf{t}; \; _{\mathbf{X}}) = P(\mathbf{y} \; j \; \mathbf{w}; \; _{\mathbf{X}}) \qquad \text{if } (\mathbf{T} \; ?? \; \mathbf{Y} \; j \; \mathbf{W}) \text{ in } G_{\mathbf{X}}: \quad (4.23)$$

Rule 2 (Change of regimes under observation):

$$P(\mathbf{y} \; j \; \mathbf{z}; \mathbf{w}; \; _{\mathbf{X}}; \; _{\mathbf{Z}}) = P(\mathbf{y} \; j \; \mathbf{z}; \mathbf{w}; \; _{\mathbf{X}}; \; ^{()}_{\mathbf{Z}}) \qquad \text{if } (\mathbf{Z} \; ?? \; \mathbf{Y} \; j \; \mathbf{W}) \text{ in } G_{\mathbf{X}\,\mathbf{Z}\underline{\mathbf{Z}}} \text{ and } G_{\mathbf{X}\,()_{\mathbf{Z}}\underline{\mathbf{Z}}}: \quad (4.24)$$

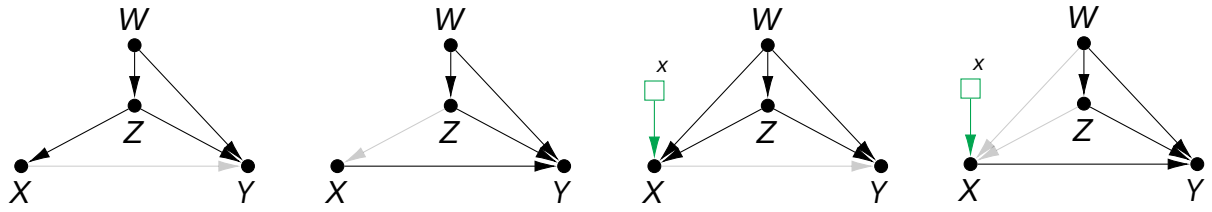
Rule 3 (Change of regimes without observation):

$$P(\mathbf{y} \; j \; \mathbf{w}; \; _{\mathbf{X}}; \; _{\mathbf{Z}}) = P(\mathbf{y} \; j \; \mathbf{w}; \; _{\mathbf{X}}; \; ^{()}_{\mathbf{Z}}) \quad \text{if } (\mathbf{Z} \; ?? \; \mathbf{Y} \; j \; \mathbf{W}) \text{ in } G_{\mathbf{X}\,\mathbf{Z}\overline{\mathbf{Z}(\mathbf{W})}} \text{ and } G_{\mathbf{X}\,()_{\mathbf{Z}}\overline{\mathbf{Z}(\mathbf{W})}}; \quad (4.25)$$

where $\mathbf{Z}(\mathbf{W}) \quad \mathbf{Z}$ *is the set of elements in* $\mathbf{Z}$ *that are not ancestors of* $\mathbf{W}$ *in* $G_{\mathbf{X}}$.

Rule 1 ascertains the validity of the *d*-separation criterion for reading conditional independence constraints in the post-interventional distribution $P(\mathbf{V}; \; _{\mathbf{X}})$, using the interventional graph $G_{\mathbf{X}}$. Rule 2 establishes a condition that guarantees that the corresponding probability distribution is the same under interventions $^{()}_{\mathbf{Z}}$ and $_{\mathbf{Z}}$ while $\mathbf{Z} = \mathbf{z}$ is observed. Rule 3 establishes a condition for changing the regime indicator from $^{()}_{\mathbf{Z}}$ to $_{\mathbf{Z}}$ without affecting the probability, whenever $\mathbf{Z}$ is not observed.

In particular, these rules can be applied with $^{()}_{\mathbf{Z}}$ having $_{Z} = ;$, for some $Z \; 2 \; \mathbf{Z}$, to make one or more regime indicators idle. When all indicators are idle, the expression corresponds to the observational regime. As noted before, differently than in the case of



(a) $G_{\underline{X}}$ (b) $G_{\overline{X}}$ (c) $G_{\ x\underline{X}}$ (d) $G_{\ x\overline{X}}$

Figure 4.3: Graphs used to test the conditions required by rules 2 and 3 of -calculus in the derivation of the query in example 18, where $_X = P(x \mid z, w)$. Arrows shown in gray indicate they have been cut.

atomic interventions and *do*-calculus, causal diagrams induced by intervened models in this context are not necessarily subgraphs of the original diagram. In practice, this means that -calculus needs to verify separation conditions in two graphs corresponding to the regimes in each side of the equality. For concreteness, we illustrate the application of the rules with the example below.

Example 20 (Tutoring program continued)**.** Let us consider again example 18 and try to identify $P(y; \ _X)$ with $_X = P(x \mid w; z)$. First, conditioning on the set $\{X, Z, W\}$ results in

$$P(y; \ _X) = \sum_{x,z,w} P(y \mid x, z, w; \ _X) P(x \mid z, w; \ _X) P(z, w; \ _X); \tag{4.26}$$

Rule 2 can be applied to the first factor with $^{()}_X = ;$ (X is playing the role of $\mathbf{Z}$ in the rule) to infer $P(y \mid x, z, w; \ _X) = P(y \mid x, z, w)$ following $(Y \perp\!\!\!\perp X \mid Z, W)$ in $G_{\underline{X}}$ and $G_{\ x\underline{X}}$ (see fig. 4.3(a) and (c), respectively). Also, Rule 3 with ($^{()}_X = ;$) leads to $P(z, w; \ _X) = P(z, w)$, licensed by the separation $(Z, W \perp\!\!\!\perp X)$ in $G_{\overline{X}}$ and $G_{\ x\overline{X}}$ (fig. 4.3(b), (d)). Then, $P(x \mid z, w; \ _X)$ is specified based on $_X = P(x \mid z, w)$, which together lead to

$$P(y; \ _X) = \sum_{x,z,w} P(y \mid x, z, w) P(x \mid z, w) P(z, w). \tag{4.27}$$

All terms on the right-hand side of eq. (4.27) are either a function of $P(\mathbf{V})$ or defined by the new intervention, which means the target effect is identifiable.

4.2.1 Comparison between -calculus and *do*-calculus

In the sequel, we list each rule in both calculi and discuss how they relate. We also underline the differences between the left and right sides of each one of the entailed constraints.

Rule 1.

$$\textit{(do-calculus)} \qquad P(\mathbf{y} \mid \mathbf{w}; \underline{\mathbf{t}}; do(\mathbf{x})) = P(\mathbf{y} \mid \mathbf{w}; \_\_; do(\mathbf{x}))$$
$$\text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{T} \mid \mathbf{W}, \mathbf{X}) \text{ in } \underline{G_{\overline{\mathbf{X}}}}: \tag{4.28}$$

$$\textit{(-calculus)} \qquad P(\mathbf{y} \mid \mathbf{w}; \underline{\mathbf{t}}; \ \mathbf{x}) = P(\mathbf{y} \mid \mathbf{w}; \_\_; \ \mathbf{x})$$
$$\text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{T} \mid \mathbf{W}) \text{ in } \underline{G_{\ \mathbf{x}}}: \tag{4.29}$$

Both rules are concerned with the validity of the d-separation criterion in an interventional setting. There are two main differences. For one, and as discussed in remark 1, due to the nature of the $do(\mathbf{X})$ intervention, there is implicit conditioning on the event $\mathbf{X} = \mathbf{x}$ hence the presence of $\mathbf{X}$ in the separation statement of eq. (4.28). In contrast, eq. (4.29) may or may not condition on $\mathbf{X}$, which depends on the set $\mathbf{W}$ or $\mathbf{T}$.

Consider the causal diagram in fig. 4.4(a) and the separation statement $(W \perp\!\!\!\perp Z)$ under $do(x)$ and $_X = P(x \mid w)$. The rules in eqs. (4.28) and (4.29) evaluate, respectively, as

$$(W \perp\!\!\!\perp Z \mid X) \text{ in } G_{\overline{X}} \text{ and} \tag{4.30}$$

$$(W \perp\!\!\!\perp Z) \text{ in } G_{\ X}: \tag{4.31}$$

While W and Z are separated in $G_{\overline{X}}$, this is not the case in $G_{\ X}$, unless X is explicitly conditioned on.

Under a soft intervention $_X$, not conditioning on X as in $P(y; \ _X)$ represents an

(a) G (b) $G_{\overline{T}}$

Figure 4.4: Model used to exemplify the use of rule 3 (see text).

average over the different values of X under the intervention, specifically

$$P(y;\ _X) = \sum_x P(y \mid x;\ _X)P(x;\ _X). \tag{4.32}$$

We can see the intervention $_X = do(x)$ represents an extreme event, because $P(x;\ _X)$ is non-zero only for $X = x$, and the average effectively disappears as it has only one element.

Another difference between these two versions of this rule, as well as rules 2 and 3, is the causal diagram where d-separation is tested. The rule in -calculus looks at $G_{\mathbf{x}}$ whereas *do*-calculus always considers $G_{\overline{\mathbf{X}}}$ which need not match.

Rule 2.

(do-calculus) $$P(\mathbf{y} \mid do(\mathbf{x}); \underline{do(\mathbf{z})}; \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}); \underline{\mathbf{z}}; \mathbf{w})$$
$$\text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}; \mathbf{W}) \text{ in } G_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}. \tag{4.33}$$

(-calculus) $$P(\mathbf{y} \mid \underset{\sim}{\mathbf{z}}; \mathbf{w};\ _{\mathbf{x}};\ _{\underline{\mathbf{z}}}) = P(\mathbf{y} \mid \underset{\sim}{\mathbf{z}}; \mathbf{w};\ _{\mathbf{x}};\ ^{0}_{\underline{\mathbf{z}}})$$
$$\text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{W}) \text{ in } G_{\ _{\mathbf{x}}\ _{\mathbf{z}}\underline{\mathbf{Z}}} \text{ and } G_{\ _{\mathbf{x}}\ ^{0}_{\mathbf{z}}\underline{\mathbf{Z}}}. \tag{4.34}$$

Both rules are related to changes in the interventional regime and observation of the intervened variables. Nevertheless, the *do*(:) operator makes the conditioning implicit in the left-hand side of eq. (4.33). For instance, consider the causal diagram in fig. 4.4(a) and the intervention $_X = do(g(w))$. The corresponding causal interventional causal diagram is

$G_{\pi_X} = G$ because X depends on the same variables in the observational and interventional regimes. Rule 2 of π-calculus guarantees that

$$P(y \mid x; w, \pi_X) = P(y \mid x; w), \tag{4.35}$$

as $(X \perp\!\!\!\perp Y \mid W)$ holds in both $G_{\underline{X}}$ and $G_{\pi_X \underline{X}}$ (same graph in this case). However, this rule cannot be used as its do-calculus counterpart, that is, exchanging π_X with $X = x$. Specifically,

$$P(y \mid w, \pi_X) = P(y \mid x; w) \tag{4.36}$$

is not entailed by the same separation statement. To witness, notice

$$P(y \mid w, \pi_X) = \sum_x P(y \mid x; w, \pi_X) P(x \mid w, \pi_X) \tag{4.37}$$

$$= \sum_x P(y \mid x; w) P(x \mid w, \pi_X), \tag{4.38}$$

where the last equality follows from eq. (4.35). In general, eq. (4.38) is not equal to $P(y \mid x; w)$.

Rule 3.

$$\textit{(do-calculus)} \quad P(\mathbf{y} \mid do(\mathbf{x}), \underline{do(\mathbf{z})}, \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \_\_, \mathbf{w})$$
$$\text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}) \text{ in } G_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}, \tag{4.39}$$

$$\textit{(π-calculus)} \quad P(\mathbf{y} \mid \mathbf{w}, \pi_{\mathbf{x}}, \pi_{\underline{\mathbf{z}}}) = P(\mathbf{y} \mid \mathbf{w}, \pi_{\mathbf{x}}, \pi^{0}_{\underline{\mathbf{z}}})$$
$$\text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{W}) \text{ in } G_{\pi_{\mathbf{x}}\, \pi_{\mathbf{z}} \overline{\mathbf{Z}(\mathbf{W})}} \text{ and } G_{\pi_{\mathbf{x}}\, \pi^{0}_{\mathbf{z}} \overline{\mathbf{Z}(\mathbf{W})}}, \tag{4.40}$$

where $\mathbf{Z}(\mathbf{W})$ $\mathbf{Z}$ is the set of elements in $\mathbf{Z}$ that are not ancestors of $\mathbf{W}$ in the corresponding graph.

Similar to the previous rules, there are differences between the diagrams and implicit

conditioning. For example, consider the causal diagram in fig. 4.4(a). Equation (4.39) together with $(Y \perp\!\!\!\perp T)$ in $G_{\overline{T}}$ (fig. 4.4(b)) leads to

$$P(y \mid do(t)) = P(y) \tag{4.41}$$

$$P(y \mid t; \pi_T = t) = P(y) \tag{4.42}$$

In contrast, for some $\pi_T = P^{*}(T \mid Y)$, the intervention cannot simply be removed from $P(y \mid t; \pi_t)$ using eq. (4.40) to obtain $P(y)$. Specifically,

$$P(y \mid t; \pi_T) = \frac{P(y; \pi_T)P(t \mid y; \pi_T)}{P(t; \pi_T)} \tag{4.43}$$

$$= P(y) \quad \frac{P(t \mid y; \pi_T)}{P(t; \pi_T)} \quad : \tag{4.44}$$

By the very definition of the intervention π_T, this is almost always different than $P(y)$. Rule 3 of π-calculus (eq. (4.40)) with $(Y \perp\!\!\!\perp T)$ in $G_{\pi_T \overline{T}}$ does entail

$$P(y; \pi_T) = P(y): \tag{4.45}$$

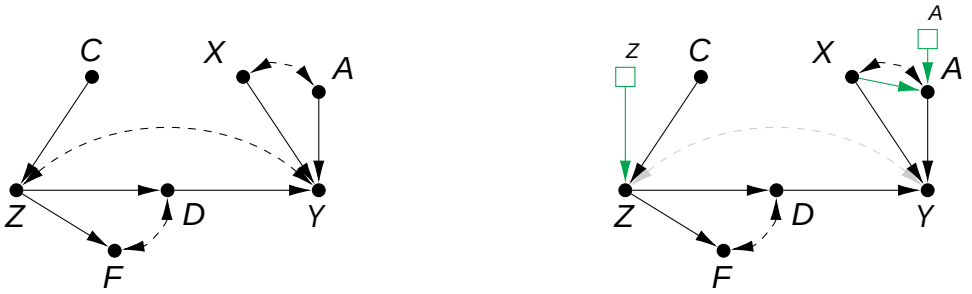
If one aims to reproduce eq. (4.41) with π-calculus when π_T is a hard intervention, two steps will be required. First, rule 1 with $(Y \perp\!\!\!\perp T)$ in $G_{\pi_T} = G_{\overline{T}}$ (fig. 4.4(b)) and then rule 3 with $(Y \perp\!\!\!\perp T)$ in $G_{\overline{T}}$, that is

$$P(y \mid do(t)) = P(y \mid t; \pi_T) \qquad \text{Def. of } \pi_T = do(t) \text{ and implicit conditioning} \tag{4.46}$$

$$= P(y; \pi_T) \qquad \text{(Rule 1)} \tag{4.47}$$

$$= P(y): \qquad \text{(Rule 3)} \tag{4.48}$$

For a more subtle example, in the context of fig. 4.4(a), let us compare the interventions $do(x); do(z)$ and $\pi_X = P^{*}(x \mid w); \pi_Z = P^{*}(z \mid x)$. For the former, rule 3 eq. (4.39) and



(a) G (b) $G_{\underline{X}\,\underline{A}}$

Figure 4.5: Pair of models associated with example 21.

$(Y \perp X \mid Z; T)$ in $G_{\overline{Z}\underline{X(T)}} = G_{\overline{Z}}$

$$P(y \mid do(x); do(z); t) = P(y \mid do(z); t). \tag{4.49}$$

However,

$$P(y \mid \underline{x}; \underline{z}; t) = P(y \mid \underline{z}; t) \tag{4.50}$$

does not hold in general for $\underline{x}$ and $\underline{z}$ as $(Y \perp\!\!\!\perp X \mid T)$ in $G_{\underline{X}\,\underline{Z}\overline{X(T)}} = G_{\underline{X}\,\underline{Z}}$ and $G_{\underline{Z}\overline{X(T)}} = G_{\underline{Z}}$.

4.2.2 A more elaborate example

We now examine a more involved setting that requires multiple applications of the -calculus.

Example 21 (A derivation in -calculus)**.** Consider the causal diagram in fig. 4.5(a) and the target conditional interventional distribution $P(y \mid x; \underline{z}; \underline{z}; \underline{a})$ with $\underline{z} = P(z|c)$ and $\underline{a} = P(a|x)$. The rules of -calculus can be used to derive this effect by first conditioning on $D; A$,

$$P(y|x; \underline{z}; \underline{z}; \underline{a}) = \sum_{d;a} P(y|x; z; d; a; \underline{z}; \underline{a}) P(d|a; x; z; \underline{z}; \underline{a}) P(a|x; z; \underline{z}; \underline{a}) \tag{4.51}$$

Let $\pi_D = do(d)$, then for the first factor in the sum:

$$P(y \mid x; z; d; a; \pi_Z; \pi_A) \tag{4.52}$$

$$= P(y \mid x; d; a; \pi_Z; \pi_A) \qquad \text{R1: } (Y \perp\!\!\!\perp Z \mid X; D; A) \text{ in } G_{\overline{Z}\,\overline{A}} \tag{4.53}$$

$$= P(y \mid x; d; a; \pi_Z) \qquad \text{R2: } (Y \perp\!\!\!\perp A \mid X; D) \text{ in } G_{\overline{Z}\,\overline{A}}, G_{\overline{Z}\underline{A}} \tag{4.54}$$

$$= P(y \mid x; d; a; \pi_Z; \pi_D) \qquad \text{R2: } (Y \perp\!\!\!\perp D \mid X; A) \text{ in } G_{\overline{Z}\underline{D}}, G_{\overline{Z}\,\overline{D}\underline{D}} \tag{4.55}$$

$$= P(y \mid x; d; a; \pi_D) \qquad \text{R3: } (Y \perp\!\!\!\perp Z \mid X; D; A) \text{ in } G_{\overline{Z}\,\overline{D}\overline{Z}}, G_{\overline{D}\overline{Z}} \tag{4.56}$$

$$= \sum_{z'} P(y \mid z'; x; d; a; \pi_D) P(z' \mid x; d; a; \pi_D) \qquad \text{condition on } Z \tag{4.57}$$

$$= \sum_{z'} P(y \mid z'; x; d; a; \pi_D) P(z'; \pi_D) \qquad \text{R1: } (Z \perp\!\!\!\perp X; D; A) \text{ in } G_{\overline{D}} \tag{4.58}$$

$$= \sum_{z'} P(y \mid z'; x; d; a; \pi_D) P(z') \qquad \text{R3: } (Z \perp\!\!\!\perp D) \text{ in } G_{\overline{D}\,\overline{D}}, G_{\overline{D}} \tag{4.59}$$

$$= \sum_{z'} P(y \mid z'; x; d; a) P(z') \qquad \text{R2: } (Y \perp\!\!\!\perp D \mid Z; X; A) \text{ in } G_{\overline{D}\underline{D}}, G_{\underline{D}} \tag{4.60}$$

Next, for the second factor

$$P(d \mid a; x; z; \pi_Z; \pi_A) \tag{4.61}$$

$$= P(d \mid z; \pi_Z; \pi_A) \qquad \text{R1: } (D \perp\!\!\!\perp A; X) \text{ in } G_{\overline{Z}\,\overline{A}} \tag{4.62}$$

$$= P(d \mid z; \pi_Z) \qquad \text{R3: } (D \perp\!\!\!\perp A \mid Z) \text{ in } G_{\overline{Z}\,\overline{A}\overline{A}}, G_{\overline{Z}\overline{A}} \tag{4.63}$$

$$= P(d \mid z) \qquad \text{R2: } (D \perp\!\!\!\perp Z) \text{ in } G_{\overline{Z}\underline{Z}}, G_{\underline{Z}}. \tag{4.64}$$

Finally, for the third factor

$$P(a \mid x, z, \sigma_Z, \sigma_A) \tag{4.65}$$

$$= P(a \mid x, \sigma_Z, \sigma_A) \qquad \text{R1: } (A \perp\!\!\!\perp Z) \text{ in } G_{\underline{Z}\,\underline{A}} \tag{4.66}$$

$$= P(a \mid x, \sigma_A) \qquad \text{R3: } (A \perp\!\!\!\perp Z) \text{ in } G_{\underline{Z}\,\underline{A}\overline{Z}}, G_{\underline{A}\overline{Z}} \tag{4.67}$$

$$= \hat{P}(a \mid x) \qquad \text{Def. of } \sigma_A\text{:} \tag{4.68}$$

Putting eqs. (4.60), (4.64) and (4.68) together leads to the following identifying expression

$$P(y \mid x, z, \sigma_Z, \sigma_A) = \sum_{d,a} \left(\sum_{z'} P(y \mid z', x, d, a) P(z') \right) P(d \mid z) \hat{P}(a \mid x). \tag{4.69}$$

Notice that since Z is observed in the query, the policy $\hat{P}(z \mid c)$, by which the value is picked, does not affect this probability.

4.3 Transportability of Soft Intervention Effects from Multiple Environments with Arbitrary Experiments from Soft Interventions[5]

This section discusses the problem of generalizing a policy $\mathbf{x}$ using assumptions encoded in a selection diagram (definition 12) from a combination of observational and experimental distributions arising from different domains.

Generalizing causal knowledge across disparate domains (i.e., populations, settings, environments) is at the heart of many inference problems across the empirical sciences, as well as AI [31, 46, 4]. In economics, for example, the Nobel Prize of 2019 was awarded to Duflo, Banerjee, and Kremer "for their experimental approach to alleviating global poverty". Their work is, in fact, about systematically performing experiments on the effect of complex policies related to poverty, and then carefully trying to extrapolate the gathered evidence to other populations [63, 64]. They acknowledge and discuss the

[5]This section is based on the paper [33].

challenges of pursuing such a task [64]: *"the number of possible variations on a given program is potentially infinite, and a theoretical framework is definitely needed to understand which variations are important to replicate and which are not."*

A typical question they try to answer could be: "If a program worked for poor rural women in Africa, will it work for middle-income urban men in South Asia?"

The same need to generalize across disparate conditions is present in Reinforcement Learning. For instance, consider a rover trained in the Californian desert for digging rocks. After exhaustive months of training, NASA wants to deploy the vehicle on Mars, where the environment is not the same, but somewhat similar to Earth. The expectation is that the rover will need minimal "experimentation" (i.e., trial-and-error) on Mars by leveraging the knowledge acquired here, operating more surgically and effectively there.

The solution to these apparently disparate tasks faced by Duflo, NASA, and so many other scientists can be framed as a transportability task (section 3.3.3). Nevertheless, instead of trying to establish the effect of a hard intervention, these problems call for the evaluation of policies entailed by soft interventions. Moreover, surrogate experiments could also be the product of soft interventions implemented in different domains.

To better understand this setting, let us consider a hypothetical study of government-backed loan programs and successful payment of family house purchases (inspired by studies such as [65, 66]).

Example 22 (Load study across cities)**.** In city , the percentage of the value of the property that can be borrowed (X) depends, mainly, on credit history (W) and current employment conditions of the applicant. The amount borrowed (or principal), Z, depends on the allowed percentage and the characteristics (R) of the property —mainly cost and location— as well as the financial characteristics of the borrower. The policy intends to increase the number of families purchasing their own homes and paying their mortgages. The outcome Y represents whether the loan is being paid promptly after a certain number of years. Figure 4.6(a) represents this situation in . Let ^{1} be another city where the distribution of

credit history W is different than in , and ^{2} another city where the distribution of R is the one that differs. In this example, the discrepancies are $^1 = fWg$ and $^2 = fRg$, which are summarized in the selection diagram G^{Δ} shown in fig. 4.6(c).

Policymakers in city are planning to increase the percentage of the property's value (X) that can be borrowed. To do so, they offer extra additional insurance for buyers of properties within certain locations (R) until the percentage left to pay falls below the usual threshold. We denote this policy as $_X$ and represent the resulting regime in fig. 4.6(b). Note that because of the new policy $_X$, it is R that points to X instead of W.

To assess the impact of this new policy, data of current payment behavior has been collected in , together with data from the other two cities, ^{1} and ^{2}. The administration of ^{1} allocated loans such that the amount Z was determined as a function of the allowed percentage (X) and the property (R). This policy is represented in fig. 4.6(d) where R points to Z and the unobservable variable encoded by W L99K Z no longer affects the decision. In city ^{2}, a study selected borrowers at random and had loaners process their applications with a randomized credit history (W), to observe the effect of W on Y. This regime is represented in fig. 4.6(e), where W does not depend on any other variable.

The effect of the new policy $_X$ can be measured by comparing $E[Y;\ _X]$ with $E[Y]$. While the latter is readily estimable via $P\ (Y)$, included in the input; the challenge is to assess $P\ (Y;\ _X)$, which we will transport later in this section (example 24).

Data available in each domain is specified by $Z = fZ^i\ j\ ^i\ 2\ g$, where each $Z^i = f\ z_1;\ z_2;:::g, \mathbf{Z}_j\ \mathbf{V}$, specifies the distributions from domain ^{i}. This means that the distributions $fP^i(\mathbf{V};\ z_j)\ j\ \mathbf{Z}_j\ 2\ Z^i g_{Z^i 2Z}$ are assumed to be available. Notice that $P^i(\mathbf{V};\ _;) = P^i(\mathbf{V})$ is a valid distribution and describes the observational (non-interventional) distribution in domain ^{i}. In example 22, $Z = fZ\ = f\ _; g; Z^1 = f\ _Z g; Z^2 = f\ _W gg$.

In general, the task to be solved in this context is

$$I\ _{-TR} = P\ (\mathbf{y};\ \mathbf{x}); \mathsf{P}; G^{\Delta}\ ; \tag{4.70}$$

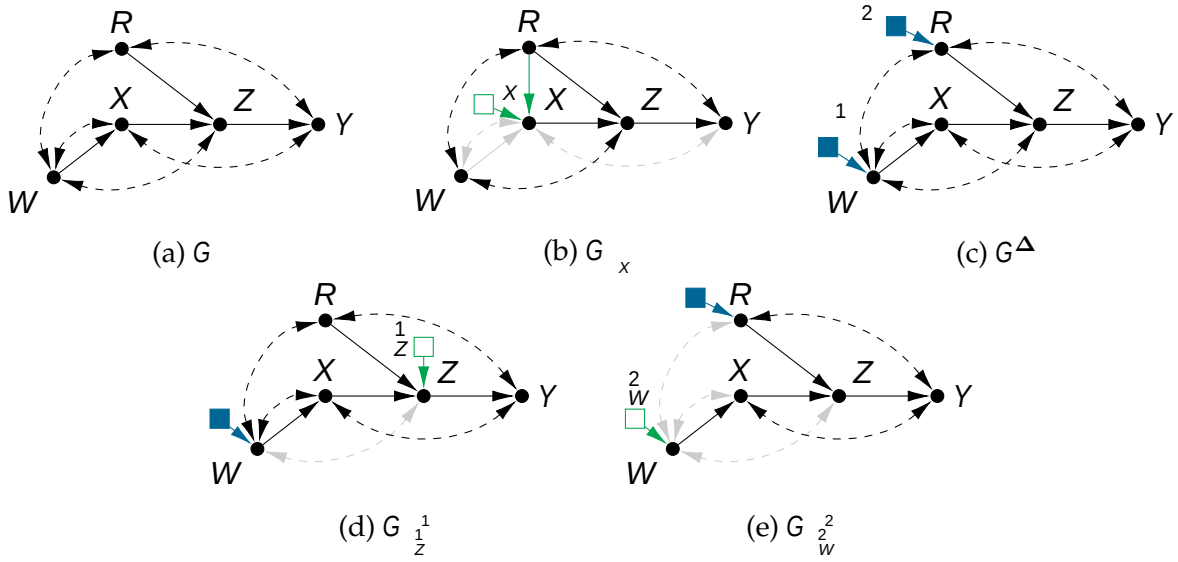


Figure 4.6: Diagram (a) represents the natural regime in and (b) the regime after the target intervention. The selection diagram (c) compares with ^{1} and ^{2}, while (d) and (e) are selection diagrams specific to domains ^{1} and ^{2} under interventions $\frac{1}{Z}$ and $\frac{2}{W}$, respectively.

where $\mathsf{P} = \{P^{i}(\mathbf{V};\ \mathbf{z}_j)\ j\ \mathbf{Z}_j\ 2\ Z^{i}\}_{Z^{i}2Z}$.

For example 22, the policy of interest can be identified from the available data using -calculus (theorem 5) and standard probability axioms. For simplicity let us write $x = {}_{X};\ z = \frac{1}{Z}$ and $w = \frac{2}{W}$ as $x;\ z$ and w. Then, the resulting expression is[6]

$$P\ (y;\ {}_{X}) = \sum_{r;x;z} \underbrace{\sum_{x'} P^{1}(yjz;x';r;\ \tfrac{1}{Z})P^{1}(x'jr;\ \tfrac{1}{Z})}_{\text{computable from } \frac{1}{Z} \text{ in } {}^{1}} \underbrace{P^{2}(zjx;r;\ \tfrac{2}{W})}_{\text{from } \frac{2}{W} \text{ in } {}^{2}} \underbrace{P\ (xjr;\ {}_{X})}_{\text{def. } {}_{X}} \underbrace{P\ (r)}_{\text{from}} : \tag{4.71}$$

Next, we will build on top of the framework introduced in chapter 3 to solve this task.

4.3.1 Solving Soft-Transportability Systematically

The -*TR* task can be systematically solved following the general strategy described in chapter 3. Specifically, C-INFER (algorithm 1) can solve I $_{\text{-}TR}$ instances with a simple generalization of c-factors (eq. (3.13)) and the -operator.

[6] See appendix E.2 for a step-by-step derivation.

Let $Q^a[\ ;\ \mathbf{x}]$ represent a c-factor defined in the context of an SCM $M^a_{\mathbf{x}}$. Then, the -operator (lemma 2) is re-defined as follows.

Lemma 4 (operator)**.** *Let* $\mathbf{T}\quad\mathbf{V}$ *be an endogenous set of variables. Then, for any cftree* T *with a node* $Q[\mathbf{T};\ \mathbf{x}]$*:*

-operator (regime invariance): If $\mathbf{T} \setminus \mathbf{X} = ;$, $Q[\mathbf{T};\ \mathbf{x}] \quad ! \quad Q[\mathbf{T};\ \overset{0}{\mathbf{x}}]$ *and* $Q[\mathbf{T};\ \mathbf{x}]$ $Q[\mathbf{T};\ \overset{0}{\mathbf{x}}]$ *are valid edges for* T *and mapped as*

$$Q[\mathbf{T};\ \overset{0}{\mathbf{x}}] = Q[\mathbf{T};\ \mathbf{x}]: \tag{4.72}$$

It follows from lemma 4 that c-factors remain invariant under soft interventions not affecting the variables in its scope. Similarly, the -operator (lemma 3) discussed in section 3.3.3 also holds when soft interventions are involved. On these considerations, GENINPUTTREE (algorithm 6) is redefined in algorithm 7.[7]

The main difference between GENINPUTTREE in algorithm 6 and algorithm 7 is the use of $\mathbf{z}$ instead of $do(\mathbf{z})$. Consequently, other operators are defined over $G_{\mathbf{z}}$ instead of $G_{\overline{\mathbf{Z}}}$. The following example is a generalization of example 19, where there is data from three domains: , ^{1} and ^{2}.

Example 23 (Generalizability of Sequential Plans)**.** The signature of the task to solve is

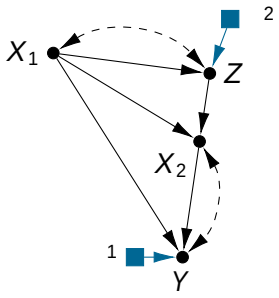
$$I = \langle P\ (y;\ _{x_1;x_2} = \ ); \mathsf{P}; G^{\Delta}\rangle; \tag{4.73}$$

where

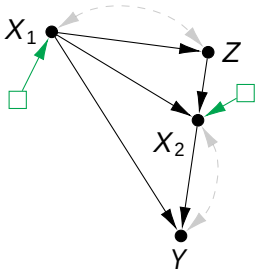
$$\mathsf{P} = \{P^1(\mathbf{V};\ _{x_1;x_2} = \{x_1; x_2\}); \qquad P^2(\mathbf{V};\ _{x_2} = g(X_1; Z)\}; \tag{4.74}$$

is $\{\ _{x_1} = P(X_1);\ \ _{x_2} = P(X_2 \mid X_1; Z)\}$ and G^{Δ} is shown in fig. 4.7(a). GENQUERYTREE

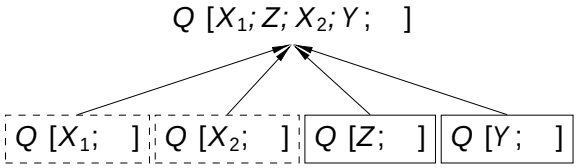
[7] Changes appear underlined.



(a) Selection diagram G^{Δ}

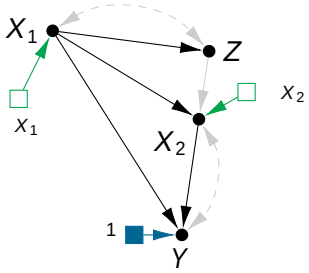


(b) G in the domain under

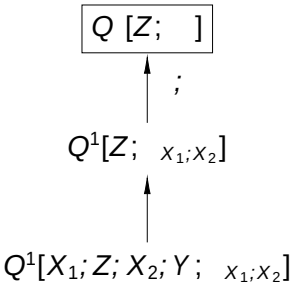


(c) $T_{P\ (y;\ )}$

1

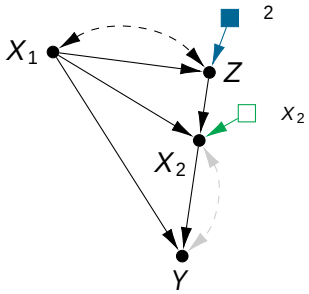


(d) G in domain ^{1} under $_{X_1;X_2} = \{x_1; x_2\}$

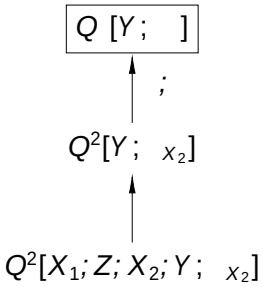


(e) $T_{P^1(\mathbf{v};\ _{X_1;X_2} = \{x_1;x_2\})}$

2



(f) G in domain ^{2} under $_{X_2} = g(X_1; Z)$



(g) $T_{P^2(\mathbf{v};\ _{X_2} = g(X_1;Z))}$

Figure 4.7: Causal diagrams and cftrees for identifying the effect of a sequential plan from heterogeneous domains.

Algorithm 7 GENINPUTTREE(P; $I = \langle Q; \mathbf{P}; G_{\mathbf{i}}; T_Q \rangle$)

Input: A distribution P, a causal inference task I and a q-tree T_Q.
Output: A d-tree T for P.

1: Initialize T with $P = Q^b[\mathbf{V}; \mathbf{x}]$ at the root.
2: **for** each node $Q^a[\mathbf{C}; \mathbf{x}] \in T_Q$, starting from the root, at every node $Q^b[\mathbf{T}; \mathbf{z}]$ **do**
3: **if** $\mathbf{C} \setminus \mathbf{Z} \neq \emptyset$; or $\mathbf{C} \setminus {}^{a;b}\!\!\not\leftarrow \neq \emptyset$; **then** give up on $Q[\mathbf{C}; \mathbf{x}]$. . Target variable intervened or different
4: **if** $\mathbf{T} = \mathbf{C}$, $\mathbf{x} = \mathbf{z}$ and $a = b$ **then** move to next $Q^a[\mathbf{C}; \mathbf{x}]$. Search is done
5: **if** $\mathbf{T} = \mathbf{C}$ and $\mathbf{C} \setminus (\mathbf{X} \cup \mathbf{Z}) = \emptyset$; **then** derive $Q^b[\mathbf{C}; \mathbf{x}]$ by σ-operator. . Model is different
6: **if** $\mathbf{T} = \mathbf{C}$ and $\mathbf{C} \setminus {}^{a;b}\!\!\not\leftarrow = \emptyset$; **then** derive $Q^a[\mathbf{C}; \mathbf{x}]$ by σ-operator. . Domain is different
7: Let $\mathbf{A} \leftarrow An(\mathbf{C})_{G_{\overline{\mathbf{z}}[\mathbf{T}]}}$.
8: **if** $\mathbf{A} \neq \mathbf{T}$ **then** use $\sum$-operator to derive and move to $Q^b[\mathbf{A}; \mathbf{z}]$. . Can sum-out variables
9: **if** $G_{\overline{\mathbf{z}}[\mathbf{T}]}$ has more than one c-component **then** use $\prod$-operator to derive $Q^b[\mathbf{W}; \mathbf{z}]$ where $\mathbf{W}$ is the union of the c-components intersecting $\mathbf{C}$. . Can factorize
10: Give up on $Q^a[\mathbf{C}; \mathbf{x}]$. . No operator left
11: **end for**

generates $T_{P_{(Y;\ )}}$ as shown in fig. 4.7(c). Notice there is no d-tree for the observational regime since $P(\mathbf{V}) \notin \mathbf{P}$. According to $T_{P_{(Y;\ )}}$, the factors needed for computing the query are $Q[X_1;\ ]$, $Q[X_2;\ ]$, $Q[Z;\ ]$ and $Q[Y;\ ]$. The first two factors are defined by the intervention $\emptyset$, therefore can be immediately mapped by MAPFACTORS as

$$Q[X_1;\ ](x_1^{\emptyset}) = 1[x_1 = x_1^{\emptyset}] \tag{4.75}$$

$$Q[X_2;\ ](x_2; x_1; z) = P^{\emptyset}(x_2 \mid x_1; z). \tag{4.76}$$

The remaining factors need to be searched for in the d-trees generated from the input distributions.

While generating $T_{P^1(\mathbf{V};\ x_1;x_2 = f(x_1;x_2))}$, GENINPUTTREE gives up (line 3) on $Q[Y;\ ]$ because $Y \in \mathbf{T}^1$ (according to the selection diagram in fig. 4.7(a)). So, it only expands towards $Q[Z;\ ]$. Similarly, when generating $T_{P^2(\mathbf{V};\ x_2 = g(X_1;Z))}$, the procedure gives up on $Q[Z;\ ]$ because $Z \in \mathbf{T}^2$ (as evidenced in fig. 4.7(d)), and only expands towards $Q[Y;\ ]$.

MAPFACTORS maps $Q[Z;\ ]$ to $T_{P^1(\mathbf{V};\ x_1;x_2)}$ and $Q[Y;\ ]$ to $T_{P^2(\mathbf{V};\ x_2 = g(X_1;Z))}$ and

COMPOSEQUERY derives expressions for these c-factors as

$$Q[Z; \quad] = Q[Z; \ _{X_1;X_2}] = Q^1[Z; \ _{X_1;X_2}] = \frac{\sum_{x_2;y} Q^1[\mathbf{V}; \ _{X_1;X_2}]}{\sum_{z;x_2;y} Q^1[\mathbf{V}; \ _{X_1;X_2}]} = P^1(z \mid x_1; \ _{X_1;X_2}); \tag{4.77}$$

$$Q[Y; \quad] = Q[Y; \ _{X_2}] = Q^2[Y; \ _{X_2}] = \frac{Q^2[\mathbf{V}; \ _{X_2}]}{\sum_{y} Q^2[\mathbf{V}; \ _{X_2}]} = P^2(y \mid x_1; x_2; z; \ _{X_2}). \tag{4.78}$$

Finally, putting eqs. (4.75) to (4.78) together, the query is given as

$$P(y; \quad) = \sum_{x_1^0;z;x_2} 1[x_1^0 = x_1] P(x_2 \mid x_1^0; z) P^1(z \mid x_1^0; \ _{X_1;X_2}) P^2(y \mid x_1^0; x_2; z; \ _{X_2}) \tag{4.79}$$

$$= \sum_{z;x_2} \underbrace{P(x_2 \mid x_1; z)}_{\text{def.}} \underbrace{P^1(z \mid x_1; \ _{X_1;X_2})}_{\text{from } P^1(\mathbf{V}; \ _{X_1;X_2}) \text{ in } \ ^1} \underbrace{P^2(y \mid x_1; x_2; z; \ _{X_2})}_{\text{from } P^2(\mathbf{V}; \ _{X_2}) \text{ in } \ ^2}. \tag{4.80}$$

The resulting expression is a combination of probabilities estimated from domain ^{1}, ^{2}, and the definition of the policies.

Next, we continue example 22 solving the task proposed in that example.

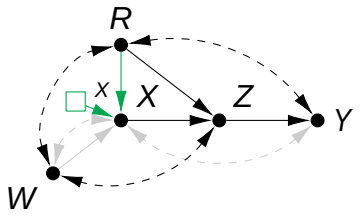
Example 24 (Loan policy generalizability — solution)**.** The signature of the task described in example 22 is

$$I = \langle P(y; \ _X); \Pi; G^{\Delta} \rangle; \tag{4.81}$$

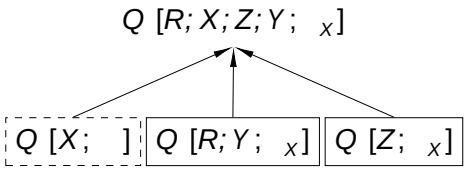
where $\Pi = \{P(\mathbf{V}); P^1(\mathbf{V}; \ _Z^1); P^2(\mathbf{V}; \ _W^2)\}$, $_X = P(X \mid R)$ and G^{Δ} is the diagram in fig. 4.8(c). In terms of c-factors, as $An(Y) = \{R; X; Z; Y\}$ in $G_{\ _X}$ (fig. 4.8(a)), the query is equal to

$$P(y; \ _X) = \sum_{r;x;z} Q[R; X; Z; Y; \ _X]. \tag{4.82}$$

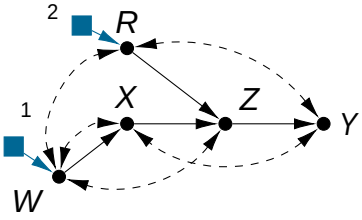
GENINPUTTREE produces $T_{P(y; \ )}$ as shown in fig. 4.8(b). There are three c-factors to obtain: $Q[X; \ _X]$, $Q[R; Y; \ _X]$ and $Q[Z; \ _X]$. The first is defined by $_X$ and will be



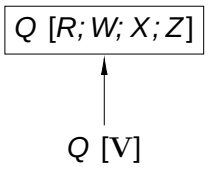
(a) G in the domain under X



(b) T_P $(y;\ )$

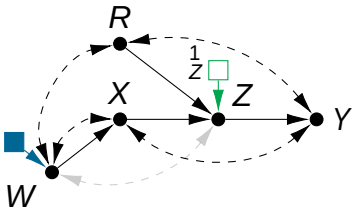


(c) Selection diagram G^{Δ}

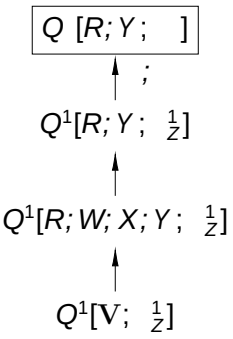


(d) T_P $(\mathbf{v})$

1

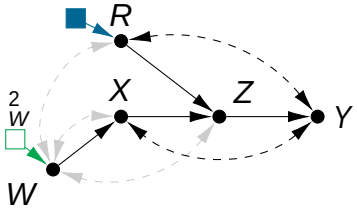


(e) G in domain ^{1} under $\frac{1}{Z}$

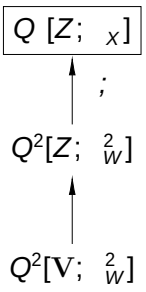


(f) $T_{P^1(\mathbf{V};\ \frac{1}{Z})}$

2



(g) G in domain ^{2} under $\frac{2}{W}$



(h) $T_{P^2(\mathbf{V};\ \frac{2}{W})}$

Figure 4.8: Causal diagrams and cftrees for transporting the effect of a soft intervention from heterogeneous domains.

mapped as

$$Q\ [X;\ \ _X] = P(x\ j\ r). \tag{4.83}$$

The factor $Q\ [R; Y;\ \ _X]$ cannot be derived in $T_{P^2(\mathbf{V};\ \frac{2}{W})=Q^2[\mathbf{V};\ \frac{2}{W}]}$ (fig. 4.8(h)) because $R\ 2\ \ ^2$. It also cannot be derived in $T_{P\ (\mathbf{V})}$ (fig. 4.8(d)), because any valid operator can only remove Y. Still, it can be found in $T_{P^1(\mathbf{V};\ \frac{1}{Z})}$ (fig. 4.8(f)). Next, $Q\ [Z;\ \ _X]$ is not computable from $P^1(\mathbf{V};\ \frac{1}{Z})$ since Z has been intervened in that model. It is also not derivable in $T_{P\ (\mathbf{V})}$ because after taking a step to marginalize Y, there are no more valid operators that retain Z. It can be, however, found in $T_{P^2(\mathbf{V};\ \frac{2}{W})}$ (fig. 4.8(h)).

The COMPOSEQUERY routine obtains expressions for the c-factors as

$$Q\ [R; Y;\ \ _X] = Q\ [R; Y;\ \ \tfrac{1}{Z}] = Q^1[R; Y;\ \ \tfrac{1}{Z}] = \sum_{w;x} Q^1[R; W; X; Y;\ \ \tfrac{1}{Z}] \tag{4.84}$$

$$= \sum_{w;x} \frac{\sum_{w;x;z;y} Q^1[\mathbf{V};\ \ \frac{1}{Z}]}{\sum_{r;w;x;z;y} Q^1[\mathbf{V};\ \ \frac{1}{Z}]} \frac{\sum_{x;z;y} Q^1[\mathbf{V};\ \ \frac{1}{Z}]}{\sum_{w;x;z;y} Q^1[\mathbf{V};\ \ \frac{1}{Z}]} \frac{\sum_{z;y} Q^1[\mathbf{V};\ \ \frac{1}{Z}]}{\sum_{x;z;y} Q^1[\mathbf{V};\ \ \frac{1}{Z}]} \frac{Q^1[\mathbf{V};\ \ \frac{1}{Z}]}{\sum_{y} Q^1[\mathbf{V};\ \ \frac{1}{Z}]} \tag{4.85}$$

$$= \sum_{w;x} P^1(w\ j\ r;\ \ \tfrac{1}{Z}) P^1(r;\ \ \tfrac{1}{Z}) P^1(x\ j\ w; r;\ \ \tfrac{1}{Z}) P^1(y\ j\ w; r; x; z;\ \ \tfrac{1}{Z}); \tag{4.86}$$

$$Q\ [Z;\ \ _X] = Q\ [Z;\ \ \tfrac{2}{W}] = Q^2[Z;\ \ \tfrac{2}{W}] = \frac{\sum_{y} Q^2[\mathbf{V};\ \ \frac{2}{W}]}{\sum_{z;y} Q^2[\mathbf{V};\ \ \frac{2}{W}]} = P^2(z\ j\ r; w; x;\ \ \tfrac{2}{W}). \tag{4.87}$$

Replacing these expression into eq. (4.82), the query $P\ (y;\ \ _X)$ is equal to

$$\sum_{r;x;z} P(xjr) \sum_{w;x'} P^1(r; w;\ \ \tfrac{1}{Z}) P^1(x' jw; r;\ \ \tfrac{1}{Z}) P^1(yjw; r; x'; z;\ \ \tfrac{1}{Z})\ \ P^2(zjr; w; x;\ \ \tfrac{2}{W}). \tag{4.88}$$

Similar to the tasks introduced in chapter 3, C-INFER together with the operators defined is sufficient and necessary for solving $\ $-*TR*, as claimed next.

Theorem 6 (Soundness and Completeness for $\ $-TR)**.** *Given a causal inference task with signature $I\ _{\text{-}TR}$, the query is transportable from* P *and* G^{Δ} *if and only if* C-INFER *finds a mapping*

using the , , and operators. Moreover, the task is decided in $O(n^2(n + m)p)$ *time, where* $n = |V|$, *m is the number of edges in G and* $p = |P|$.

Moreover, the sequence of operators entailed by the paths of the cftrees generated by C-INFER on a -*TR* task can be mapped to a derivation with -calculus and probability axioms, like those given in section 4.2.2. Then, from theorem 6 we have the following:

Corollary 1 (-calculus Completeness for -*TR*)**.** *The -calculus together with standard probability axioms is complete for the task of -TR.*

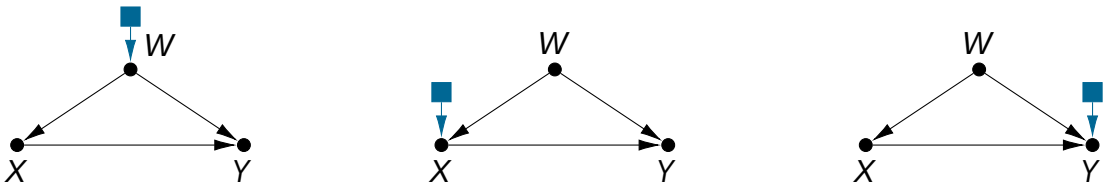
Chapter 5: Statistical Transportability

Generalizing causal and statistical findings across settings is central in scientific inference as well as in many applications throughout artificial intelligence and machine learning. The environment where the data is collected (source) is related to, but not often the same as the one where the inferences are intended (target). No learning could occur if the target environment is arbitrary or drastically different from the training (source) environment. However, the fact that we learn and perform relatively well in new environments suggests that certain characteristics are shared across environments and that, owing to these commonalities, causal and statistical claims could be valid and robust even in settings where no or very little data is available [31, 46, 22, 4].

Remarkably, the anchors of knowledge that allow extrapolations to take place are eminently causal, following from the stability of the mechanisms shared across settings [67]. The systematic analysis of these mechanisms and the conditions under which extrapolation could be formally justified fall under the umbrella of transportability theory.

Despite all the progress achieved so far, most of these results focused on the conditions under which *interventional distributions* could be extrapolated, leaving a class of generalization problems with no solution, namely, noninterventional distributions. In practice, on the other hand, many problems in AI and ML today, including classification and clustering, entail the learning of a (non-interventional) probability distribution of the form $P(\mathbf{y} \mid \mathbf{x})$, where, for example, $\mathbf{Y}$ could represent a label and $\mathbf{X}$ a set of features.[1] Depending on the differences between environments, the distribution $P(\mathbf{y} \mid \mathbf{x})$ may or may not be a good predictor in the target domain. The mismatch between source and target environments is

[1] Note that, in contrast to the challenges in the previous chapters, this inference problem is trivially solved if $P(\mathbf{V})$ is available.



(a) (b) (c)

Figure 5.1: Selection diagrams representing three different generalization tasks.

due to the differences in the underlying causal mechanisms.

This generalization problem has been studied in machine learning as *dataset shift* [19], *transfer learning* [15], and *domain adaptation* [16]. Moreover, it has been studied from a causal inference perspective [17, 18], and called *statistical transportability* by Pearl and Bareinboim [68]. The statistical transportability problem has been formalized but not solved in a systematic and nonparametric version. For concreteness, consider the following example.

Example 25 (Extrapolating demographic information)**.** An internet company records in its database the type of advertisement displayed to its customers (X) and whether the corresponding product was sold (Y) on its website . The user's age (W) affects both the ad format (X) as well as her/his propensity for buying the product (Y). The distribution $P(x; w; y)$ can be estimated with great accuracy from this large dataset.

The company plans to expand to a different country and call the data science team to help predict Y given X, $P(yjx)$, in the new market. The populations and could differ in different ways. For instance, the following scenarios exemplify some differences between source and target domains, which lead to different analyses and conclusions:

Scenario 1 (S1) the age distribution is significantly different ($P(W) \neq P(W)$),

Scenario 2 (S2) the website is run by another team and uses a different strategy for selecting the ad format ($P(X \mid W) \neq P(X \mid W)$), and

Scenario 3 (S3) the buying behavior (Y) differs, for instance, since users are less wealthy in ($P(Y) \neq P(Y)$).

These differences can be explained by variations of the underlying mechanism across

settings. In other words, each domain has a different SCM. As for the previous transportability tasks, differences are encoded with selection diagrams and selection variables (definition 12). The diagrams in fig. 5.1 correspond to the three scenarios described above.

For (S1), the causal analyst in the team suggests that they should use data from the census of population , and estimate the new age distribution, $P^{*}(w)$. The team goes on to say that this will allow the target query $P^{*}(y \mid x)$ to be written in a convenient form:

$$P^{*}(y \mid x) = \frac{\sum_{w} P(y \mid x, w) P(x \mid w) P^{*}(w)}{\sum_{w} P(x \mid w) P^{*}(w)}. \tag{5.1}$$

The right-hand side of the expression is estimable by combining data from the source ($P(\mathbf{V})$) and the smaller dataset from the target ($P^{*}(W)$).

For (S2), it suffices to just observe the strategy for selecting the ad format in , $P^{*}(X \mid W)$, noting that the target query can be written as

$$P^{*}(y \mid x) = \sum_{w} P(y \mid x, w) \frac{P^{*}(x \mid w) P(w)}{\sum_{w} P^{*}(x \mid w) P(w)}. \tag{5.2}$$

Finally, for (S3), $P^{*}(y \mid x)$ needs to be re-learned from scratch with data from .

It is worth noting that even if the target quantity is a conditional distribution of Y, it could be estimated without measuring Y in the target domain. This is indeed the case in example 25 for the first and second scenario as eqs. (5.1) and (5.2) only require $P^{*}(W)$ and $P^{*}(X \mid W)$ to be measured in the target domain, respectively.

The main goal of this chapter is to explicate the rationale behind this analysis and, more broadly, to provide a systematic way of deciding statistical transportability. Formally, the signature of the task we aim to solve is

$$I_{s\text{-}TR} = \langle P^{*}(\mathbf{y} \mid \mathbf{x}); \mathbf{P}; G^{\Delta} \rangle; \tag{5.3}$$

where $\mathbf{P} = \{P(\mathbf{V}); P^{*}(\mathbf{W})\}$, $\mathbf{W}$ $\mathbf{V}$, and G^{Δ} is a selection diagram.

Whenever there exists a mapping $P(\mathbf{V})$ and $P(\mathbf{W})$ to $Q = P(y \mid x)$ for set of models inducing G^{Δ}, Q will be called *statistically transportable* (or simply *transportable*).

Compared to the tasks discussed in chapters 3 and 4, the *s-TR* task presents two main differences. First, the query is an observational distribution instead of a causal one. Second, one of the distributions is marginal over a subset of the observable (endogenous) variables. Notice that if $\mathbf{W} = \mathbf{V}$ or $(\mathbf{X} \cup \mathbf{Y}) \subseteq \mathbf{W}$, the problem would be trivial since $P(y \mid x)$ would be immediately computable from $P(\mathbf{W})$.

The statistical transportability tasks has connections with machine learning research, specifically with problems related to (observational) predictions. Many results in the field are concerned with learning good representations of a probability distribution of the form $P(y \mid x)$, where Y is the label or category to be predicted and $\mathbf{X}$ is a set of features used for prediction. On this front, neural networks have led to great advances due to their ability to approximate virtually any function [69, 70, 71]. Other issues arise due to the complexity of the model, as well as the number of samples required to learn it efficiently. Nevertheless, even if those issues were solved and a perfect model for $P(y \mid x)$ could be learned from the available data, the model could perform poorly if the underlying data generating process differs in some mechanism with the process in the target/test domain [17].

Pearl and Bareinboim [68] formally defined the task in a non-parametric setting as a variation of transportability where the target is an observational quantity instead of a causal effect. It is assumed that the causal diagram is known, $P(\mathbf{V})$ is available in the source domain, and $P(\mathbf{W})$ is available in the target domain for some $\mathbf{W} \subseteq \mathbf{V}$ ($\mathbf{W}$ could be empty). This task was named *statistical transportability* and [68] illustrated the solution of some instances of the problem. In this chapter we consider the statistical transportability task, introducing first some sufficient graphical conditions to solve special instances (section 5.1). Later on, we introduce a generalization of c-factors (section 5.2) and a canonical way to rewrite conditional queries (section 5.3), leading to a systematic approach capable of solving any instance of this task (section 5.4).

5.1 Sufficient Graphical Conditions

The assumptions encoded in the selection diagram play a critical role in understanding what needs to be measured in the target domain. To witness, consider the different diagrams in fig. 5.2 over three variables $X; Z$ and Y. The figure is organized in a way such that all models in the same row share the same graphical structure and those in the same column share the same structural invariances. For instance, fig. 5.2(e) has a "common cause" structure where Z 2 , that is, there could be differences in f_z or $P(\mathbf{U}_z)$ between and . The other two graphs in the same row share the same structure, but in fig. 5.2(g) the differences are due to f_x or $P(\mathbf{U}_x)$, whereas for fig. 5.2(i) they are due to f_y or $P(\mathbf{U}_y)$. Furthermore, selection diagrams in the same column share the same discrepancies between the source and target domain where fig. 5.2(b) has a chain structure and fig. 5.2(h) has a collider structure. By and large, each selection diagram entails different requirements for the data needed from domain , as described in the corresponding captions.

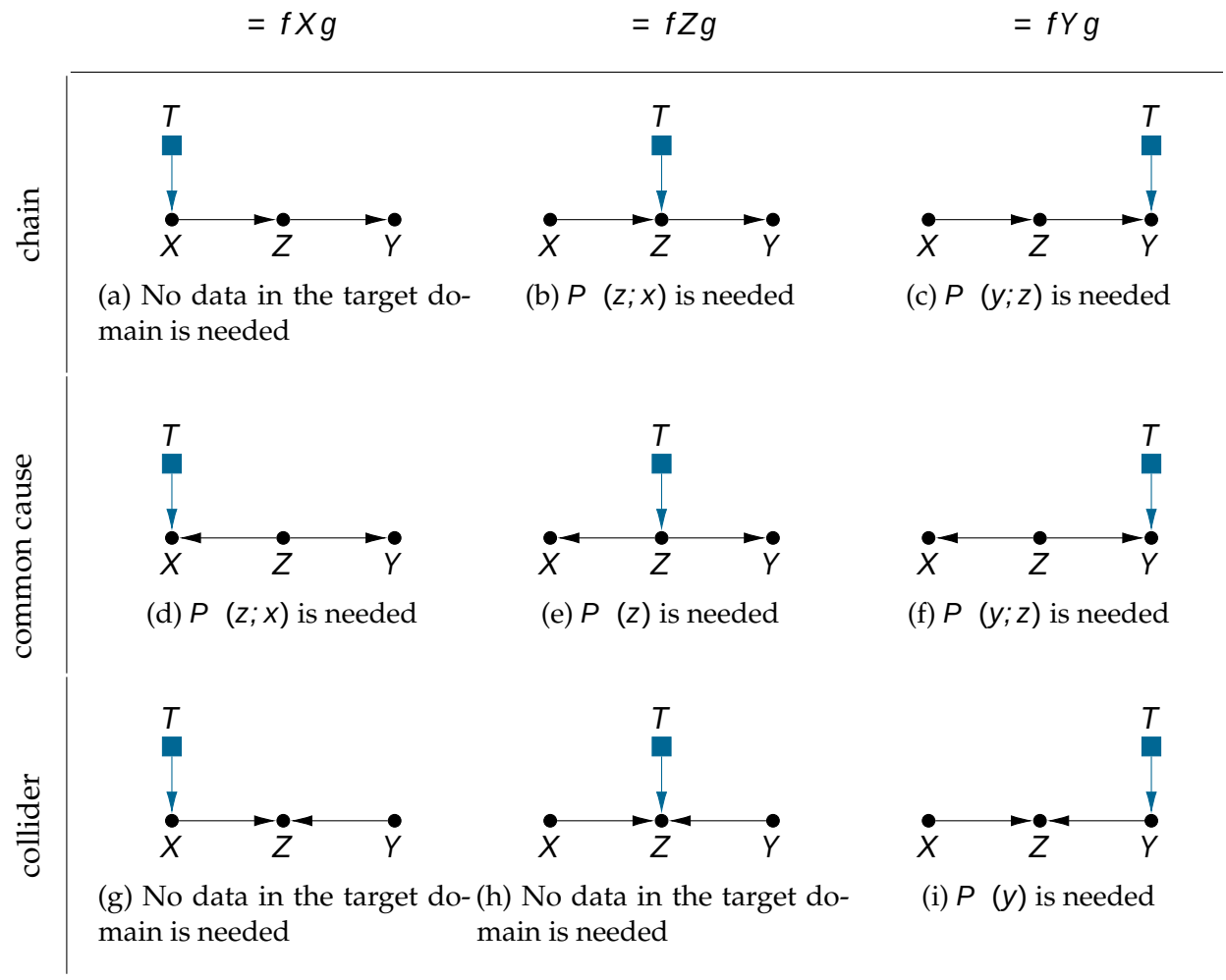
The variables that we need to measure in are not the same across rows or columns. This aspect of the problem is not completely determined by the structure or the differences between the domains. Instead, such analysis requires careful consideration of the interplay between those aspects of the models. This suggests that the knowledge provided by the selection diagram (or equivalent assumptions) is necessary for solving the s-*TR* task.

To determine if a (conditional) distribution is the same in the source and target domain based on the selection diagram, the d-separation criterion[2] can be used for selection variables, denoted $\mathbf{T}$.[3] Specifically,

$$P(\mathrm{y} \mid \mathrm{x}) = P\ (\mathrm{y} \mid \mathrm{x}) \qquad \text{if } (\mathbf{T} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}). \tag{5.4}$$

[2]The d-separation criterion allows one to read conditional independence statements in the distribution by looking at paths in the graphical model. See appendix A.1 for a definition of the criterion.

[3]These variables were originally named S by Bareinboim and Pearl [23] but here we rename them as T, to avoid conflict with chapter 6 where S is used to represent the sampling mechanism.



(a) No data in the target domain is needed

(b) P $(z; x)$ is needed

(c) P $(y; z)$ is needed

(d) P $(z; x)$ is needed

(e) P (z) is needed

(f) P $(y; z)$ is needed

(g) No data in the target domain is needed

(h) No data in the target domain is needed

(i) P (y) is needed

Figure 5.2: Variations of a causal diagram made of only a path with three variables, with differences between source and target domain at different variables. For each causal diagram the question is what variables need to be observed in the target domain for transporting P $(y \mid x)$.

In this sense, selection variables can be thought of as a switch between the domains and . If $\mathbf{T}$ is d-separated from $\mathbf{Y}$ given $\mathbf{X}$, then $\mathbf{Y}$ is not affected by the differences across these domains and remains invariant to the changes in mechanism that took place across domains.

The rest of this section discusses some conditions and examples for the statistical transportability of the query $P(\mathbf{y} \mid \mathbf{x})$ from the input in $I_{s\text{-}TR}$, and the selection diagram G^{Δ}.

Proposition 1 (No measurements collected in). *Assume no data is collected in the target domain , so* $\mathbf{W} = ;$. *Then* $P(\mathbf{y} \mid \mathbf{x})$ *is transportable from* $P(\mathbf{V})$ *if and only if* $(\mathbf{Y} \perp\!\!\!\perp \mathbf{T} \mid \mathbf{X})$ *in which case it is equal to*

$$P(\mathbf{y} \mid \mathbf{x}) = P(\mathbf{y} \mid \mathbf{x}). \tag{5.5}$$

This result characterizes situations where no data is available in the target domain. For instance, models in fig. 5.2(a), (g), and (h) could be solved via proposition 1.

When some data is available in the target domain, a first case to consider is when the measurements include the features in the query.

Proposition 2 (The set of features $\mathbf{X}$ is measured in). *Assume data from a set of covariates used in the query,* $\mathbf{X}$, *is measured in the target domain , that is* $\mathbf{X} \quad \mathbf{W}$. *Then* $P(\mathbf{y} \mid \mathbf{x})$ *is transportable from* $P(\mathbf{V})$ *and* $P(\mathbf{W})$ *if there exists a set* $\mathbf{Z} \quad \mathbf{W} \cap \mathbf{X}$ *(possibly empty) such that* $(\mathbf{T} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}, \mathbf{Z})$ *and, if so, the query distribution in is equal to*

$$P(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z} \mid \mathbf{x}). \tag{5.6}$$

The models (b) and (d) from fig. 5.2 can be solved if X and Z are measured in .

Instead of the set of features $\mathbf{X}$, it could be the outcome $\mathbf{Y}$ that is measured in , as covered in the following condition.

Proposition 3 (The outcome $\mathbf{Y}$ is measured in). *Assume the outcome $\mathbf{Y}$ is measured in the target domain , that is, $\mathbf{Y}$ $\mathbf{W}$. Then $P(\mathrm{y} \mid \mathrm{x})$ is transportable from $P(\mathbf{V})$ and $P(\mathbf{W})$ if there exists a set $\mathbf{Z}$ $\mathbf{W}$ n $\mathbf{Y}$ (possibly empty) such that $(\mathbf{T}; \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z})$ and given as*

$$P(\mathrm{y} \mid \mathrm{x}) = \frac{\sum_{\mathrm{z}} P(\mathrm{y} \mid \mathrm{z}) P(\mathrm{x} \mid \mathrm{z}) P(\mathrm{z})}{\sum_{\mathrm{z}} P(\mathrm{x} \mid \mathrm{z}) P(\mathrm{z})}. \tag{5.7}$$

In this case, it is fig. 5.2(c), (f) and (i) that could be solved with proposition 3.

One may surmise, given these conditions, that either the set of features $\mathbf{X}$ or the outcome $\mathbf{Y}$ need to be measured in the target domain . The following result shows this is not the case.

Proposition 4 (Neither $\mathbf{X}$ nor $\mathbf{Y}$ are measured in). *If there exists $\mathbf{Z}$ $\mathbf{W}$ such that $(\mathbf{Y}; \mathbf{X} \perp\!\!\!\perp \mathbf{T} \mid \mathbf{Z})$, then $P(\mathrm{y} \mid \mathrm{x})$ is transportable from $P(\mathbf{V})$ and $P(\mathbf{W})$ and given as*

$$P(\mathrm{y} \mid \mathrm{x}) = \frac{\sum_{\mathrm{z}} P(\mathrm{y} \mid \mathrm{x}; \mathrm{z}) P(\mathrm{x} \mid \mathrm{z}) P(\mathrm{z})}{\sum_{\mathrm{z}} P(\mathrm{x} \mid \mathrm{z}) P(\mathrm{z})}. \tag{5.8}$$

The selection diagram in fig. 5.2(e) describes a model where proposition 4 applies with $\mathbf{Z} = \mathbf{W} = \{Z\}$.

To illustrate how to use these graphical conditions and evaluate them, we consider some additional examples.

Example 26 (Graphical conditions for statistical transportability). Let us consider each of the selection diagrams in fig. 5.3 in the context of a task $I_{s\text{-}TR}$ with different sets $\mathbf{W}$.

First, for the selection diagram in fig. 5.3(a) and $\mathbf{W} = ;$, the query is transportable by proposition 1. Specifically, the separation statement $(T \perp\!\!\!\perp Y \mid X)$ holds in the diagram, so:

$$P^{*}(y \mid x) = P(y \mid x). \tag{5.9}$$

This means that even though the mechanism for X could differ between and , by conditioning on X those disparities do not affect Y.

The same condition is not valid for the selection diagram in fig. 5.3(b) since the path $! X \leftarrow Z ! Y$ is d-connected given X. Instead, if the set $\mathbf{W} = \{X, Z\}$ is measured, proposition 2 can be used as $(Y \perp\!\!\!\perp T \mid X, Z)$ holds. In this case,

$$P^{*}(y \mid x) = \sum_{z} P(y \mid x, z) P^{*}(z \mid x). \tag{5.10}$$

The situation in fig. 5.3(c) is particularly stringent as there is a selection variable pointing directly to Y, which implies that without measuring Y itself in the target domain (i.e., $Y \in \mathbf{W}$), generalization from source to the target domain is provable infeasible. Now, suppose $\mathbf{W} = \{Y, Z\}$, then, by proposition 3 and $(X \perp\!\!\!\perp T, Y \mid Z)$ it follows that

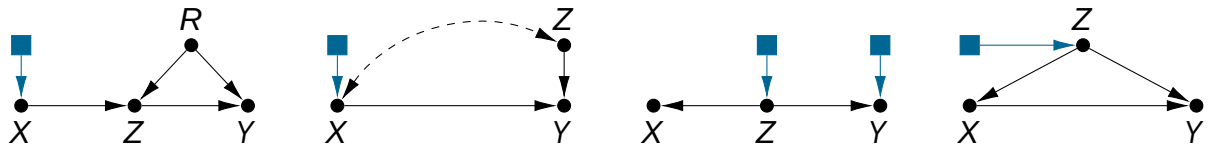
$$P^{*}(y \mid x) = \frac{\sum_{z} P(y \mid z) P(x \mid z) P^{*}(z)}{\sum_{z} P(x \mid z) P^{*}(z)}. \tag{5.11}$$

Finally, suppose $\mathbf{W} = \{Z\}$ is measured in for fig. 5.3(d). In this case, proposition 4 is the only valid option since $(X, Y \perp\!\!\!\perp T \mid Z)$, which means the target distribution can be written as

$$P^{*}(y \mid x) = \frac{\sum_{z} P^{*}(y \mid x, z) P(x \mid z) P^{*}(z)}{\sum_{z} P(x \mid z) P^{*}(z)}. \tag{5.12}$$

In other words, neither the features $\mathbf{X}$ nor the outcome $\mathbf{Y}$ need to be measured in the target domain, just the other features that changed (Z).

Those conditions cover a wide range of instances of the problem, however, the interplay between the graphical structure and the discrepancies between domains may require a



(a) (b) (c) (d)

Figure 5.3: Some examples of selection diagrams and s-TR tasks can be solved with simple graphical conditions.

more surgical approach to the problem. The following sections build the tools for an algorithmic analysis of the problem.

5.2 Marginalized c-factors and c-components

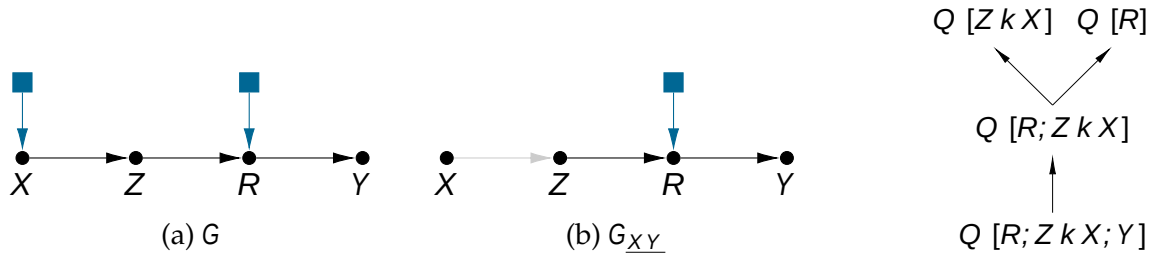
Despite the general (and somewhat intuitive) nature of the results in the previous section, there are still cases in which the target distribution is transportable and propositions 1 to 4 are unable to say so. For instance, if we measure $\mathbf{W} = \{Z; R\}$ in settings following fig. 5.4(a), the query could be transported as

$$P(y \mid x) = \sum_{z;w} P(y \mid r)P(r \mid z)P(z \mid x); \tag{5.13}$$

which is not obtainable from any of the previous conditions.

This section develops refined concepts for solving the s-TR task in generality, using the algorithmic approach introduced in chapter 3. One of the main differences between this task and the previous ones is that *not all* the variables in $\mathbf{V}$ are observed for some of the input distributions. Namely, the set of variables $\mathbf{W}$ observed in the target domain is a strict subset of $\mathbf{V}$. If only the target domain was considered, treating the variables in $\mathbf{V} \cap \mathbf{W}$ as unobservables could be a valid choice, yet this approach loses information in the causal diagram and the measures on $\mathbf{V} \cap \mathbf{W}$ in other domains.

This new dimension requires several nontrivial changes in the underlying machinery, including a generalization of the notion of c-factor (definition 8), which should take into



(a) G (b) $G_{\underline{XY}}$ (c) d-tree for $P(R;Z)$

Figure 5.4: The marginalized c-components of the selection diagram in (a) summing $\{Z;R\}$ are the connected components in the subgraph in (b). The d-tree in fig. 5.4(c) is based on the same selection diagram and has marginalized c-factor for nodes.

account marginalized variables.[4]

Definition 14 (Marginalized c-factor)**.** For disjoint subsets $\mathbf{C};\mathbf{H} \subseteq \mathbf{V}$, $Q[\mathbf{C} \,k\, \mathbf{H}]$ is the *marginalized c-factor* of $\mathbf{C}$ where variables in $\mathbf{H}$ are marginalized out and denotes the following function

$$Q[\mathbf{C} \,k\, \mathbf{H}](pa(\mathbf{c} \mid \mathbf{h}) \cap \mathbf{h}) = \sum_{\mathbf{h}} Q[\mathbf{C};\mathbf{H}]. \tag{5.14}$$

Definition 15 (Marginalized c-components)**.** Let G be a causal diagram over $\mathbf{V}$ and let $\mathbf{H} \subseteq \mathbf{V}$. Two variables $V_i; V_j \in \mathbf{V}$ belong to the same *c-component, summing* $\mathbf{H}$ if there exists a path (regardless of the directionality) between V_i and V_j in $G_{\underline{\mathbf{V} \cap \mathbf{H}}}$.[5]

Whenever $\mathbf{H} = \emptyset$, $G_{\underline{\mathbf{V} \cap \mathbf{H}}} = G_{\underline{\mathbf{V}}}$, which is the same as G with all directed edges removed. Since $G_{\underline{\mathbf{V}}}$ has only the bidirected edges from G, the marginalized c-components summing $\emptyset$ are simply the c-components of G.

Example 27 (Marginalized c-factors and c-components)**.** Consider again the graph in

[4]The very definition of SCM already assumes certain variables are not measured, called exogenous. The subtlety is that in this setting certain variables are measured in one domain but not in another, so they cannot be classified as exogenous.

[5]When a set $\mathbf{H}$ is marginalized, a c-factor over $\mathbf{C}$ and the c-components of some G are referred to as "the c-factor of $\mathbf{C}$ summing $\mathbf{H}$" and "the c-components of G summing $\mathbf{H}$".

fig. 5.4(a) and the query $Q = P(y \mid x)$. We first note that the query can be written as

$$P(y \mid x) = \frac{P(y, x)}{\sum_y P(y, x)}. \tag{5.15}$$

We further write $P(y, x)$ in terms of the joint distribution $P(\mathbf{V})$ by summing over $\mathbf{V} \setminus \{X, Y\} = \{Z, R\}$. Then,

$$P(y, x) = \sum_{r,z} P(x, z, r, y) \tag{5.16}$$

$$= \sum_{r,z} Q[X, Z, R, Y] = Q[X, Y \cup Z, R]. \tag{5.17}$$

Moreover, G has two c-components summing $\{Z, R\}$: $\{X\}$ and $\{Z, R, Y\}$. This is easy to verify visually by looking at the graph $G_{\underline{XY}}$ in fig. 5.4(b). Interestingly, these marginalized c-components advertise a way to factorize $P(x, y)$ as follows.

$$P(x, z, r, y) = P(x)P(z \mid x)P(r \mid z)P(y \mid r) \tag{5.18}$$

$$\sum_{r,z} P(x, z, r, y) = \sum_{r,z} P(x)P(z \mid x)P(r \mid z)P(y \mid r) \tag{5.19}$$

$$P(x, y) = \underbrace{P(x)}_{Q[X]} \underbrace{\sum_{z,r} P(z \mid x)P(r \mid z)P(y \mid r)}_{Q[Y \cup Z, R]}. \tag{5.20}$$

That is, an expression can be factorized according to its marginalized c-components when sums are involved. This resembles the $\prod$-operator that licenses factorization according to the c-component structure.

As a generalization of c-factors, marginalized c-factors can also serve as a basis for generating causal factor trees. All cftree operators introduced so far naturally extend to the marginalized case, shown next.

Theorem 7 ($\sum$; $\prod$; and operators — marginalized c-factors)**.** *Let* $\mathbf{T}, \mathbf{L} \subseteq \mathbf{V}$ *be disjoint sets*

of endogenous variables, $\mathbf{C}$ $\mathbf{T}$, $\mathbf{C}^{0} = \mathbf{T} \cap \mathbf{C}$, $\mathbf{H}$ $\mathbf{L}$ *and* $\mathbf{H}^{0} = \mathbf{L} \cap \mathbf{H}$. *Then, for any cftree T*

-operator (marginalization): If there is no directed arrow with tail in $\mathbf{C}^{0} \cup \mathbf{H}^{0}$ *and head in* $\mathbf{C} \cup \mathbf{H}$ *in* $G[\mathbf{T}; \mathbf{L}]$, $Q[\mathbf{T} \setminus \mathbf{L}]$! $Q[\mathbf{C} \setminus \mathbf{H}]$ *is a valid edge for T associated with the mapping*

$$Q[\mathbf{C} \setminus \mathbf{H}] = \sum_{\mathbf{c}^{0}} Q[\mathbf{T} \setminus \mathbf{L}] \tag{5.21}$$

-operator (independence): If there is no path between $\mathbf{C} \cup \mathbf{H}$ *and* $\mathbf{C}^{0} \cup \mathbf{H}^{0}$ *in* $G[\mathbf{T}; \mathbf{L}]_{\underline{\mathbf{T}}}$, $Q[\mathbf{T} \setminus \mathbf{L}]$! $Q[\mathbf{C} \setminus \mathbf{H}]$ *is a valid edge for T . The associated function is defined for any topological order* $T_1 < T_2 <$ $< T_k$ *on* $G[\mathbf{T}; \mathbf{L}]$ *as follows*

$$Q[\mathbf{C} \setminus \mathbf{H}] = \prod_{T_i \in \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \setminus \mathbf{L}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \setminus \mathbf{L}]}. \tag{5.22}$$

Furthermore, let $(\mathbf{C}_1; \mathbf{H}_1), \ldots, (\mathbf{C}_k; \mathbf{H}_k)$ *be the marginalized c-components of* $G[\mathbf{T}]$, *summing* $\mathbf{L}$, *then the set of edges* $\{Q[\mathbf{T} \setminus \mathbf{L}]$ $Q[\mathbf{C}_i \setminus \mathbf{H}_i]\}_{i=1,\ldots,k}$ *are valid for T and the corresponding functional mapping is*

$$Q[\mathbf{T} \setminus \mathbf{L}] = \prod_{i=1}^{k} Q[\mathbf{C}_i \setminus \mathbf{H}_i]. \tag{5.23}$$

-operator (regime invariance): If $(\mathbf{T} \cup \mathbf{L}) \setminus \mathbf{X} = \emptyset$, $Q[\mathbf{T} \setminus \mathbf{L}; \mathbf{x}]$! $Q[\mathbf{T} \setminus \mathbf{L}; \mathbf{x}^{0}]$ *and* $Q[\mathbf{T} \setminus \mathbf{L}; \mathbf{x}]$ $Q[\mathbf{T} \setminus \mathbf{L}; \mathbf{x}^{0}]$ *are valid edges for T and they are mapped as*

$$Q[\mathbf{T} \setminus \mathbf{L}; \mathbf{x}^{0}] = Q[\mathbf{T} \setminus \mathbf{L}; \mathbf{x}]. \tag{5.24}$$

-operator (domain invariance): If $(\mathbf{T} \cup \mathbf{L}) \setminus$ $^{a,b} = \emptyset$, $Q^{a}[\mathbf{T} \setminus \mathbf{L}]$! $Q^{b}[\mathbf{T} \setminus \mathbf{L}]$ *and* $Q^{a}[\mathbf{T} \setminus \mathbf{L}]$ $Q^{b}[\mathbf{T} \setminus \mathbf{L}]$ *are valid edges for T and they are mapped as*

$$Q^{b}[\mathbf{T} \setminus \mathbf{L}] = Q^{a}[\mathbf{T} \setminus \mathbf{L}]. \tag{5.25}$$

Example 28 (cftree operators on marginalized c-factors)**.** Consider again the selection diagram in fig. 5.4(a) and the distribution $P\ (Z; R)$ given as data in the target domain. This distribution can be written as

$$P\ (z; r) = \sum_{x;y} P\ (x; z; r; y) \tag{5.26}$$

$$= \sum_{x;y} Q\ [X; Z; R; Y] = Q\ [Z; R \,k\, X; Y]: \tag{5.27}$$

The d-tree in fig. 5.4(c) is rooted at this c-factor and illustrates some of the operators that can be used on top of it. First, the -operator allows deriving $Q\ [Z; R \,k\, X]$ as there is no arrow from Y to any other node in G. Then, there are no paths (regardless of directionality) between fRg and $fZ; Xg$ in $G[X; Z; R]_{\underline{ZR}}$, hence the -operator licenses the expansion of $Q\ [Z; R \,k\, X]$ to $Q\ [R]$ and $Q\ [Z \,k\, X]$.

Although there is no reason to move away from the observational regime in the target domain for this task, the and operators are also valid in this case. For instance, $Q\ [Z \,k\, X] = Q[Z \,k\, X]$ because $fZ; Xg \setminus \ = \ ;$, that is, there is no selection variable pointing to $fZ; Xg$.

Marginalized c-factors and c-components will enable the construction of d-trees involving marginal distributions given as input. In the next section, they will be used to characterize the c-factors needed to evaluate a conditional distribution.

5.3 Conditional Query Factorization

A target conditional query entails a ratio between two marginal distributions, by definition. Interestingly, some cancellations may occur if those marginal distributions decompose into more than one factor. For concreteness, let us see an example where the factorization entailed by the -operator (eq. (5.23)) can help simplify the c-factors needed for mapping a conditional query.

Example 29 (Factor cancellation in a query)**.** Recall the selection diagram in fig. 5.4(a) and the query $Q = P(y \mid x)$ which is equal to the ratio $P(y; x)/P(x)$. From eq. (5.17), the joint distribution $P(y; x) = Q[X; Y \,k\, Z; R]$. The goal is to determine if $P(y; x)$ and $P(x) = \sum_y P(y; x)$ have common c-factors that could cancel out. Based on the -operator:

$$P(y; x) = Q[X; Y \,k\, Z; R] = Q[X]Q[Y \,k\, Z; R]: \tag{5.28}$$

$$P(x) = \sum_y Q[X; Y \,k\, Z; R] = Q[X \,k\, Z; R; Y] = Q[X]: \tag{5.29}$$

Here $Q[X \,k\, Z; R; Y] = Q[X]$ follows from the -operator and the fact that none of $Z; R$, and Y have arrows pointing to X in G.

Suppose only Z and R are measured in , that is, let the set of input distributions be $\mathsf{P} = \{P(X; Z; R; Y); P(Z; R)\}$. Then, the c-factor $Q[Y \,k\, Z; R] = \sum_{z;r} Q[Z]Q[R]Q[Y]$ can be evaluated from P since by the -operator $Q[Z] = Q[Z]$ and $Q[Y] = Q[Y]$ which are computable from $P(\mathbf{V})$. In addition, $Q[R]$ is computable from $P(Z; R) = Q[Z; R \,k\, X]$ via -operator as shown in the d-tree in fig. 5.4(c).

Meanwhile, $Q[X] = P(x)$ cannot be transported from P. This is because it may be different than $P(x)$ due to the selection variable pointing to X in G_{Δ} and because X is not measured in $P(Z; R)$. Nevertheless, since the c-factor $Q[X]$ appears both in the numerator and denominator, it can be canceled out so that

$$P(y \mid x) = Q[Y \,k\, Z; R]: \tag{5.30}$$

In other words, although there are two factors related to $P(y; x)$, the conditional query $P(y \mid x)$ depends only on $Q[Y \,k\, Z; R]$. This means that, even if $Q[X]$ is not transportable, the transportability of the query $P(y \mid x)$ is dictated by that of $Q[Y \,k\, Z; R]$.

The strategy used in this example can be generalized by rewriting any conditional

query as

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{P(\mathbf{y}, \mathbf{x})}{P(\mathbf{x})} \qquad \text{(cond. probability def.)} \qquad (5.31)$$

$$= \frac{\sum_{\mathbf{v} \in (\mathbf{y}, \mathbf{x})} Q[\mathbf{V}]}{\sum_{\mathbf{v} \in \mathbf{x}} Q[\mathbf{V}]} \qquad (P(\mathbf{v}) = Q[\mathbf{V}]) \qquad (5.32)$$

$$= \frac{\sum_{\mathbf{d} \in (\mathbf{y}, \mathbf{x})} Q[\mathbf{D}]}{\sum_{\mathbf{d} \in \mathbf{x}} Q[\mathbf{D}]} \qquad \text{(by } \sigma\text{-operator)} \qquad (5.33)$$

$$= \frac{\sum_{\mathbf{y}} Q[\mathbf{Y}, \mathbf{X} \cup \mathbf{D} \cap (\mathbf{Y}, \mathbf{X})]}{\sum_{\mathbf{y}} Q[\mathbf{Y}, \mathbf{X} \cup \mathbf{D} \cap (\mathbf{Y}, \mathbf{X})]}; \qquad \text{(definition 14)} \qquad (5.34)$$

where $\mathbf{D} = An(\mathbf{Y}, \mathbf{X})$.

Once the query is written as eq. (5.34), the question is which (marginal) c-factors are actually needed for transportability to succeed. To address this issue using q-trees, a conditional notion of c-factor is provided below.

Definition 16 (Conditional c-factor)**.** For disjoint subsets $\mathbf{C}_1, \mathbf{C}_2 \subseteq \mathbf{V}$, $Q[\mathbf{C}_1 \mid \mathbf{C}_2]$ is said to be the *conditional c-factor* of $\mathbf{C}_1$ given $\mathbf{C}_2$ if

$$Q[\mathbf{C}_1 \mid \mathbf{C}_2](pa(\mathbf{c}_1, \mathbf{c}_2)) = \frac{Q[\mathbf{C}_1, \mathbf{C}_2]}{\sum_{\mathbf{c}_1} Q[\mathbf{C}_1, \mathbf{C}_2]}. \qquad (5.35)$$

Following from eq. (5.34), any query $P(\mathbf{y} \mid \mathbf{x})$ can be written as

$$P(\mathbf{y} \mid \mathbf{x}) = Q[\mathbf{Y} \mid \mathbf{X} \cup \mathbf{D} \cap (\mathbf{Y}, \mathbf{X})], \text{ where } \mathbf{D} = An(\mathbf{Y}, \mathbf{X}). \qquad (5.36)$$

For instance, for the selection diagram in fig. 5.5(a), the query $P(y \mid x)$ is equal to

$$P(y \mid x) = Q[Y \mid X \cup Z, R]. \qquad (5.37)$$

The following tree operator allows for reducing the scope of a conditional c-factor under some conditions.

Lemma 5 (-operator)**.** *Let* $\mathbf{T}; \mathbf{L}$ $\mathbf{V}$ *be disjoint sets,* $\mathbf{C}$ $\mathbf{T}$, $\mathbf{C}^{0} = \mathbf{T} \cap \mathbf{C}$, $\mathbf{H}$ $\mathbf{L}$ *and* $\mathbf{H}^{0} = \mathbf{L} \cap \mathbf{H}$*. Then, for any cftree* T *:*

-operator (unconditioning): Let $\mathbf{T}_1; \mathbf{T}_2$ $\mathbf{T}$ *be disjoint sets such that* $\mathbf{T}_1 \cup \mathbf{T}_2 = \mathbf{T}$*. If* $\mathbf{T}_1$ $\mathbf{C}$ $\mathbf{T}$ *and there is no path between* $\mathbf{C} \cup \mathbf{H}$ *and* $\mathbf{C}^{0} \cup \mathbf{H}^{0}$ *in* $G[\mathbf{T}; \mathbf{L}]_{\underline{\mathbf{T}_2}}$, $Q[\mathbf{T}_1 \mid \mathbf{T}_2 \cup \mathbf{L}]$ $Q[\mathbf{C} \cup \mathbf{H}]$ *is a valid edge for* T *. The associated function is*

$$Q[\mathbf{T}_1 \mid \mathbf{T}_2 \cup \mathbf{L}] = \frac{\sum_{\mathbf{t}_1} Q[\mathbf{C} \cup \mathbf{H}]}{\sum_{\mathbf{t}_1} Q[\mathbf{C} \cup \mathbf{H}]} \tag{5.38}$$

The -operator establishes a connection between conditional c-factors and marginalized c-factors beyond eq. (5.35). While the definition of conditional c-factor gives

$$Q[\mathbf{T}_1 \mid \mathbf{T}_2 \cup \mathbf{L}] = \frac{\sum_{\mathbf{t}_1} Q[\mathbf{T}_1; \mathbf{T}_2 \cup \mathbf{L}]}{\sum_{\mathbf{t}_1} Q[\mathbf{T}_1; \mathbf{T}_2 \cup \mathbf{L}]}; \tag{5.39}$$

the c-factor $Q[\mathbf{C} \cup \mathbf{H}]$ on the right-hand-side of eq. (5.38) could have a smaller scope than $Q[\mathbf{T}_1; \mathbf{T}_2 \cup \mathbf{L}]$, depending on the graphical structure.

In the following example, the -operator is used to find the c-factors associated with the query (those that do not cancel out).

Example 30 (Relevant query c-factors)**.** Consider the selection diagram in fig. 5.5(a) and the query $P(y \mid x)$. To use the -operator, first notice that for $\mathbf{D} = An(X; Y) = \{R; Z; X; A; Y\}$, then by eq. (5.36) it follows

$$P(y \mid x) = Q[Y \mid X \cup R; Z; A]: \tag{5.40}$$

Then, consider the graph $G[\mathbf{D}]$ in fig. 5.5(b) and its marginalized c-components summing X. These c-components are precisely the connected components of the graph $G[\mathbf{D}]_{\underline{X}}$, namely, $\{X; R; Z\}$ and $\{Y; A\}$. The only marginalized c-component intersecting Y is $\{Y; A\}$, hence there is no path in $G[\mathbf{D}]_{\underline{X}}$ between $\{Y; A\}$ and any other variables in $\mathbf{D}$. Applying the

(a) G (b) $G[An(X;Y)]_{\underline{X}}$

Figure 5.5: A selection diagram and the subgraph are used to determine the relevant (marginalized) c-factors for the target query.

-operator with $\mathbf{T}_1 = fYg$, $\mathbf{T}_2 = fXg$, and $\mathbf{L} = fR;Z;Ag$ entails

$$P(y \mid x) = \frac{Q[Y \mid kA]}{\sum_y Q[Y \mid kA]} = Q[Y \mid kA] = \sum_a P(y \mid x;a)P(a); \tag{5.41}$$

as there are no paths from $fY;Ag = \mathbf{C} \cup \mathbf{H}$ to $fX;R;Zg = \mathbf{C}^{0} \cup \mathbf{H}^{0}$ in $G[fY;X;R;Z;Ag]_{\underline{X}}$ (fig. 5.5(b)).

In other words, this means that the query is not affected by the differences in $fZ;Rg$ between the target and source domains.

5.4 Solving Statistical Transportability Algorithmically[6]

The algorithmic framework introduced in chapter 3 needs to be augmented to handle input distributions where only a subset of the endogenous variables is measured.

Depending on what variables are observed for a certain input distribution, the factorization of the query may change accordingly. Algorithm 8 generalizes GENQUERYTREE in algorithm 2, used for the previous tasks.

Compared to algorithm 2, this version uses the -operator to reduce the query following the discussion in section 5.3, expanding $Q[\mathbf{A} \mid \mathbf{H}]$. Moreover, this version expands alternative subtrees in T_Q for distinct subsets $\mathbf{W}$ $\mathbf{V}$ in the input distributions (line 5). For each subtree, the variables in $\mathbf{H}$ not included in $\mathbf{W}$ remain marginalized as $Q[\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W}) \mid \mathbf{H} \cap \mathbf{W}]$. Then, each (marginalized) c-component of $Q[\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W}) \mid \mathbf{H} \cap \mathbf{W}]$,

[6]This section is based on the paper [34].

Algorithm 8 GENQUERYTREE($I = \langle Q; \mathcal{P}; \mathcal{G} \rangle$)

Input: A causal inference task such that $Q = P(\mathbf{y} \mid \mathbf{x})$ ($\mathbf{x}$ could be empty) and there exists $\mathcal{G} \in \mathcal{G}$ describing the SCM associated with Q. (Q could be interventional in which case $\mathcal{G}$ represents an intervention graph.)
Output: T_Q, an q-tree for Q.

1: Let $\mathcal{G} \in \mathcal{G}$ be the causal diagram associated with the query.
2: Let $\mathbf{D} \leftarrow An(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}}}$.
3: Initialize T_Q with Q at the root.
4: Expand $Q = Q[\mathbf{Y} \mid \mathbf{X}] \leftarrow Q[\mathbf{A} \mid \mathbf{H}]$ where $(\mathbf{A}, \mathbf{H})$ is the union of the marginalized c-components of $\mathcal{G}[\mathbf{D}]$, summing $\mathbf{D} \setminus \mathbf{X}$, that intersects with $\mathbf{Y}$.
5: **for each** distinct $\mathbf{W} \subseteq \mathbf{V}$ such that some $P \in \mathcal{P}$ is over $\mathbf{W}$ **do**
6: Expand $Q[\mathbf{A} \mid \mathbf{H}] \leftarrow Q[\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W}) \mid \mathbf{H} \cap \mathbf{W}]$.
7: Expand $Q[\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W}) \mid \mathbf{H} \cap \mathbf{W}] \leftarrow Q[\mathbf{A}_i \mid \mathbf{H}_i]$ for each marginalized c-component of $\mathcal{G}[\mathbf{A} \cup \mathbf{H}]_{\overline{\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W})}}$.
8: Expand each $Q[\mathbf{A}_i \mid \mathbf{H}_i] \leftarrow Q[\mathbf{A}_i, \mathbf{H}_i]$
9: Expand each $Q[\mathbf{A}_i, \mathbf{H}_i] \leftarrow Q[\mathbf{C}_i]$ for each c-component of $\mathcal{G}[\mathbf{A}_i \cup \mathbf{H}_i]$.
10: **end for**
11: **return** T_Q

denoted as $Q[\mathbf{A}_i \mid \mathbf{H}_i]$, is expanded (line 7) as they could match c-factors of $P(\mathbf{W})$. Lines 8 and 9 further expand non-marginalized c-factors $Q[\mathbf{A}_i, \mathbf{H}_i]$ and decompose them in smaller $Q[\mathbf{C}_i]$, in case they are computable individually.

Notice that for q-trees supporting marginalized and conditional c-factors, the root of T_Q is the query itself (instead of $Q[\mathbf{D}]$). Then, another small change needed for this task is that the last line of COMPOSEQUERY (algorithm 4) just needs to return the expression for the root of T_Q, instead of computing a sum.

On the input side, any distribution $P(\mathbf{W})$, $\mathbf{W} \subseteq \mathbf{V}$, can be written as

$$P(\mathbf{w}) = \sum_{\mathbf{v} \cap \mathbf{w}} P(\mathbf{v}) \tag{5.42}$$

$$= \sum_{\mathbf{a} \cap \mathbf{w}} P(\mathbf{a}); \tag{5.43}$$

where $\mathbf{A} = An(\mathbf{W})$. Hence, as illustrated in example 31, the d-tree for such $P(\mathbf{W})$ can be generated by GENINPUTTREE (algorithm 7) with the root $P(\mathbf{w}) = Q[\mathbf{W} \mid An(\mathbf{W}) \cap \mathbf{W}]$.

Example 31 (Solving statistical transportability systematically)**.** Consider the task $I = \langle P(y \mid x); \mathsf{P}; G^{\Delta}\rangle$, where $\mathsf{P} = \{P(\mathbf{V}); P(X; Z)\}$ and G^{Δ} is given in fig. 5.6(a).

First, GENQUERYTREE looks at the c-components of $G[An(Y; X)]$ (fig. 5.6(b)) summing X, and produces the tree T_Q shown in the top portion of fig. 5.6(d). While the subtree rooted at $Q[R; Z; X; C; A; Y]$ is generated for $\mathbf{W} = \mathbf{V}$, the one rooted at $Q[Z; X; Y \mid R; C; A]$ is generated for $\mathbf{W} = \{Z; X\}$ (lines 5-9).

Next, GENINPUTTREE (algorithm 7) generates a d-tree $T_{P(\mathbf{V})}$ for $P(\mathbf{V})$ as shown in the bottom-left section of fig. 5.6(d). Meanwhile, distribution $P(\mathbf{W}) = P(Z; X)$ corresponds to the marginalized c-factor $Q[Z; X \mid R]$. To witness, first write it as a function of $P(\mathbf{V})$:

$$P(z; x) = \sum_{\mathbf{v} \sim \{z;x\}} P(\mathbf{v}) = \sum_{r} P(r; z; x) = Q[Z; X \mid R]. \tag{5.44}$$

Hence, $T_{P(Z;X)}$ can also be generated by GENINPUTTREE with $Q[Z; X \mid R]$ at the root and using marginalized c-components with the corresponding operators.

The process succeeds as the query can be connected to the inputs via the c-factors $Q[Z; X \mid R]$, $Q[A; Y]$ and $Q[C]$. Finally, the corresponding mapping is obtained by composing the functions associated with the edges in the tree. The mapping obtained is

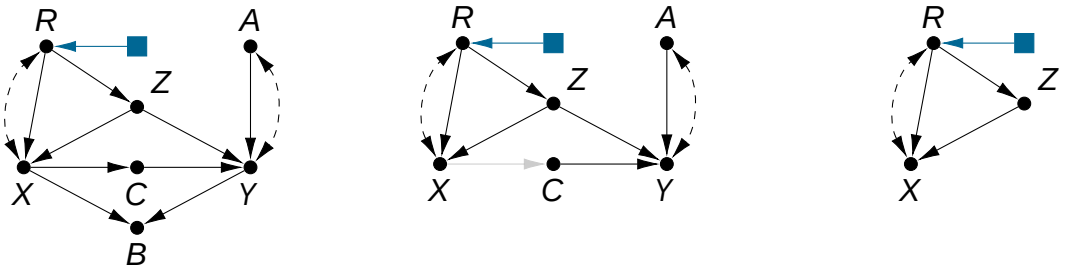
$$P(y \mid x) = \frac{\sum_{z} Q[Z; X \mid R] \sum_{a;c} Q[A; Y] Q[C]}{\sum_{y;z} Q[Z; X \mid R] \sum_{a;c} Q[A; Y] Q[C]} \tag{5.45}$$

$$= \frac{\sum_{z} P(z; x) \sum_{a;c} P(a) P(y \mid r; z; x; c; a) P(c \mid r; z; x)}{\sum_{y;z} P(z; x) \sum_{a;c} P(a) P(y \mid r; z; x; c; a) P(c \mid r; z; x)} \tag{5.46}$$

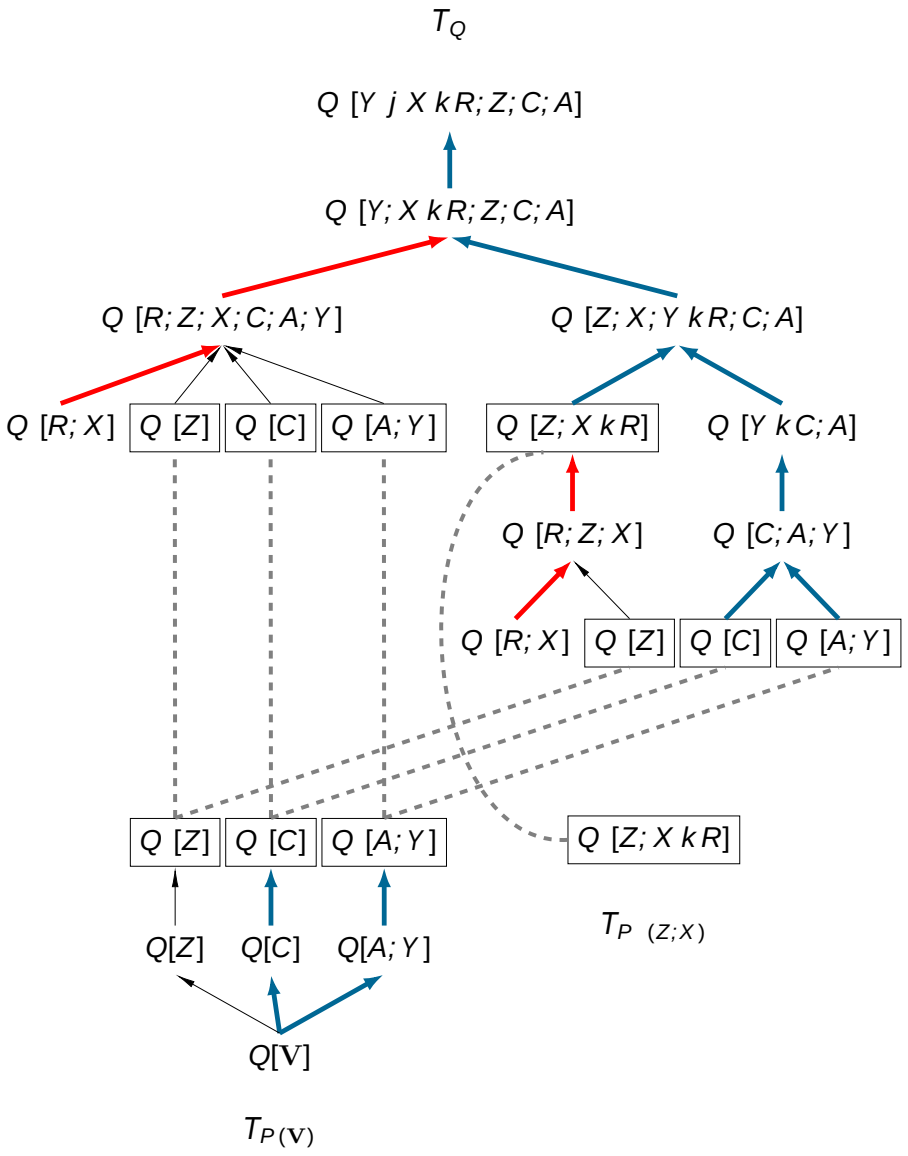
$$= \sum_{z} P(z \mid x) \sum_{a;c} P(a) P(y \mid r; z; x; c; a) P(c \mid r; z; x). \tag{5.47}$$

As shown next, the C-INFER strategy is complete for $I_{s\text{-}TR}$.

Theorem 8 (, , and -operators soundness and completeness for *s-TR*)**.** *Given a causal inference task with signature $I_{s\text{-}TR}$, the query Q is transportable from P and G^{Δ} if and only if* C-INFER *finds a mapping using the , , and -operators. Moreover, the task is decided in $O(n^2(n + m))$ time, where $n = |\mathbf{V}|$ and m is the number of edges in G.*



(a) G (b) $G[An(Y;X)]_{\underline{X}}$ (c) $G[An(\mathbf{W})] = G[An(Z;X)]$



(d) Representation of the matching process of the query and d-trees

Figure 5.6: Selection diagrams associated with the input distributions and query of an s-TR task. For each one of the distributions a d-tree is generated.

Chapter 6: Recovering from Selection Bias

Selection bias arises when there is preferential inclusion of the subjects in the data sample [22]. For instance, in a typical study of the effect of grades on college admission, subjects with higher achievement tend to report their scores more frequently than those who scored lower. In this case, the data-gathering process will reflect a distortion in the sample's proportions and, since the data is no longer a faithful representation of the underlying population, biased estimates will be produced regardless of the number of samples collected (even if the treatment is controlled).

The problem of selection bias can also be modeled graphically through the explicit articulation of the sampling mechanism, S. This mechanism can be seen as a binary indicator of entry into the data pool, such that $S = 1$ if a unit is included in the sample and $S = 0$ otherwise. When the sampling process is entirely random, S is independent of all variables in the analysis. When samples are collected preferentially, the causal effects need to be identified and *recovered* from the distribution $P(\mathbf{V} \mid S = 1)$ instead of $P(\mathbf{V})$ [72].

Selection bias has challenged inferences throughout a wide range of disciplines, including AI [73, 74, 75, 76], statistics [77, 78, 79, 80, 81], and the empirical sciences (e.g., genetics [82, 83], economics [84, 85], and epidemiology [86, 87]).

Even though selection and confounding biases appear together in most of the non-trivial, practical settings, they have been almost invariably treated independently in the literature. There are non-trivial interactions between them, however, which just recently have been investigated. [26, 1] provided sufficient conditions for the nonparametric recoverability of the causal effects from selection bias and introduced a relaxation of this setting so that external (unbiased) data could be leveraged. [81] developed an approach for discrete models, where assumptions on the cardinality of the observable variables allow to

estimate the distribution over the sampling mechanism; in turn, recovering the marginal distribution.

While most tasks introduced so far are concerned with determining a causal effect from observational, experimental, and multiple domain distributions, this chapter enriches the algorithmic framework developed in this dissertation by allowing for selection biased distributions to be taken as input.

Section 6.1 introduces the basic notions found in the literature including covariate adjustment and related results. In section 6.2, we introduce a sufficient and necessary graphical condition for recovering causal effects via adjustment [31, p. 78] from a selection-biased observational distribution and a nonbiased distribution over a set of covariates. Later on, section 6.3 presents another sufficient and necessary adjustment criterion, this time for generalizing selection-biased experimental distributions to some target domain. Finally, section 6.4 extends the algorithmic framework presented so far to include support distributions of the form $P(\mathbf{V} \mid S = 1)$ as input.

6.1 Covariate Adjustment

Before discussing the results in sections 6.2 and 6.3, we present some concepts and results related to covariate adjustment that will be necessary for later discussion. We begin with the most simple version of an adjustment criterion — the *Backdoor Criterion*.

Definition 17 (Adjustment [31])**.** Given a causal diagram G containing a set of variables $\mathbf{V}$ and pairwise disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, the set $\mathbf{Z}$ is called covariate adjustment for estimating the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ (or usually just adjustment), if for every distribution $P(\mathbf{v})$ compatible with G it holds that

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z}). \tag{6.1}$$

Finding an adjustment set relative to $\mathbf{X}$ and $\mathbf{Y}$ enables the identification of the corresponding causal effect. Several criteria have been developed to determine whether a set $\mathbf{Z}$ is valid for adjustment. The most representative result for controlling for confounding bias by adjustment is known as the *Backdoor criterion* [47, 31], which is defined as follows:

Definition 18 (Backdoor Criterion)**.** A set of variables $\mathbf{Z}$ satisfies the Backdoor Criterion relative to a pair of variables $(X; Y)$ in a directed acyclic graph G if:

(i) No node in $\mathbf{Z}$ is a descendant of X.

(ii) $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X.

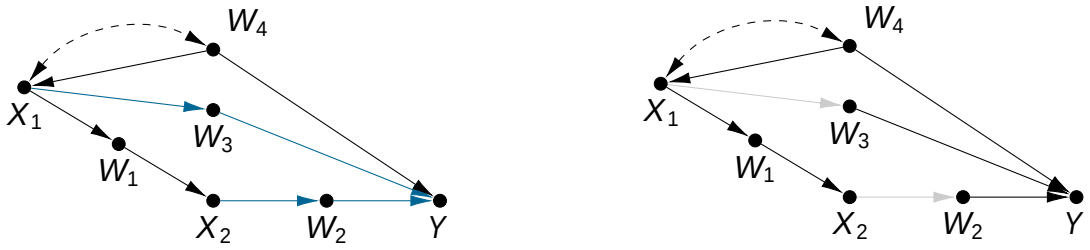
Intuitively, the backdoor criterion identifies the sets that block the non-causal paths (paths with arrows incoming towards X) while leaving the causal paths undisturbed.

It was further noted that certain descendants of X could be included in the adjustment set without sacrificing its validity [88]. Moreover, it can be shown that an adjustment is valid if and only if there exists some $\mathbf{Z}$ that blocks all paths but for the proper causal ones [89], as defined below.

Definition 19 (Proper Causal Path [89])**.** Let $\mathbf{X}$ and $\mathbf{Y}$ be sets of nodes in a causal diagram. A causal path from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$ is called proper if it does not intersect $\mathbf{X}$ except at the starting point.

Example 32 (Proper causal paths and proper back-door graph)**.** Consider the causal diagram in fig. 6.1(a) and let $X_1; X_2$ be the treatments and Y be the outcome. There are three causal paths: $X_1 \to W_1 \to X_2 \to W_2 \to Y$, $X_1 \to W_3 \to Y$, and $X_2 \to W_2 \to Y$. However, only the last two causal paths are proper since the first one is intersected by X_2.

The following definition describes the construction of a graph where only the non-causal and non-proper causal paths remain, which will facilitate the search of a set of covariates $\mathbf{Z}$ that block the appropriate paths.



(a) G (b) $G^{pbd}_{\mathbf{XY}}$

Figure 6.1: (a) A causal diagram with proper causal paths from $X_1; X_2$ to Y highlighted. (b) A proper back-door graph relative to $X_1; X_2$ and Y.

Definition 20 (Proper Back-door Graph [90])**.** Let G be a causal diagram, and $\mathbf{X}; \mathbf{Y}$ be disjoint subsets of variables. The proper backdoor graph relative to $\mathbf{X}$ and $\mathbf{Y}$, denoted as $G^{pbd}_{\mathbf{XY}}$, is obtained from G by removing the first edge of every proper causal path from $\mathbf{X}$ to $\mathbf{Y}$.

The causal diagram in fig. 6.1(b) is the $G^{pbd}_{\mathbf{XY}}$ of fig. 6.1(a) relative to $X_1; X_2$ and Y. Following the definition, the first edge of the proper causal paths $X_1 \to W_3 \to Y$, and $X_2 \to W_2 \to Y$ have been removed. All remaining paths between $\mathbf{X} = \{X_1; X_2\}$ and Y are spurious and would need to be blocked to satisfy the conditions for adjustment.

6.2 Recovering Causal Effects by Adjustment[1]

Bareinboim, Tian, and Pearl [26] pointed out that adjustment can be used for controlling for selection bias, in addition to confounding. They then proposed a sufficient graphical condition called *Selection-Backdoor* criterion. For instance, for the model in fig. 6.2(a), the criterion allows us to write the effect of X on Y as a function of the biased distribution $P(\mathbf{V} \mid S = 1)$ and non-biased data $P(Z)$, as follows:

$$P(y \mid do(x)) = \sum_z P(y \mid x; z; S = 1)P(z): \tag{6.2}$$

Although the selection backdoor is sufficient, it is not necessary and does not capture

[1]This section is based on the papers [35, 36].

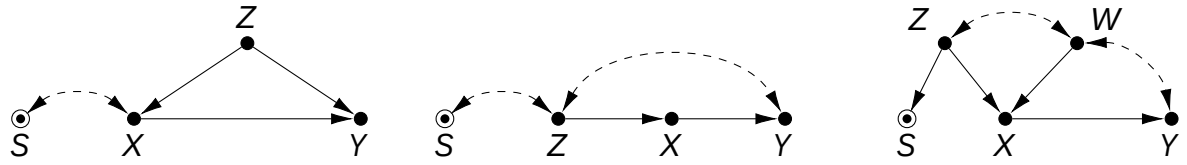


Figure 6.2: Causal diagrams with a selection bias node. For models (a) and (b) the causal effect $P(y \mid do(x))$ can be recovered by adjustment from $P(\mathbf{v} \mid S = 1)$. For the model in (c) the same effect can be recovered by a mapping other than adjustment.

other adjustment expressions. For instance, they further noted that for a model with the diagram in fig. 6.2(b), the same causal effect can be expressed as a function of the biased distribution, as:

$$P(y \mid do(x)) = \sum_{z} P(y \mid x; z; S = 1) P(z \mid S = 1). \tag{6.3}$$

There are also models for which a causal effect cannot be expressed in terms of an adjustment expression but can be written as a different function of the biased distribution $P(\mathbf{V} \mid S = 1)$. For instance, there is no adjustment functional for the effect of X on Y in fig. 6.2(c), yet this effect is recoverable as

$$P(y \mid do(x)) = \sum_{w} P(y \mid x; w; z; S = 1) P(w \mid z; S = 1). \tag{6.4}$$

In this section, a sufficient and necessary adjustment criterion for recovering causal effects from biased data together with unbiased data measured over a subset of the observed variables. Formally, the signature of the task is

$$I_{sbt\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})), \mathsf{P}, G_S \rangle, \tag{6.5}$$

where $\mathsf{P} = \{P(\mathbf{V} \mid S = 1), P(\mathbf{T})\}$ and $\mathbf{T} \quad \mathbf{V}$.

For this particular section, the question is whether Q can be expressed as eq. (6.6) (below) for some set $\mathbf{Z} \quad \mathbf{V}$ $n(\mathbf{X} \cup \mathbf{Y})$ and $\mathbf{Z}^{\mathbf{T}} \quad \mathbf{Z} \setminus \mathbf{T}$. The following definition formalizes

the notion of a valid pair of such sets.

Definition 21 (Adjustment Pair)**.** Given a causal diagram G augmented with selection variable S, disjoint sets of variables $\mathbf{X}; \mathbf{Y}; \mathbf{Z}$, and a set $\mathbf{Z}^{\mathbf{T}}$ $\mathbf{Z}$, $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ is said to be an *adjustment pair* for recovering the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ if for every model compatible with G it holds that:

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}; \mathbf{z}; S = 1) P(\mathbf{z} \cap \mathbf{z}^{\mathbf{T}} \mid \mathbf{z}^{\mathbf{T}}; S = 1) P(\mathbf{z}^{\mathbf{T}}) \tag{6.6}$$

To illustrate this definition, note that if the adjustment pair is $(Z; Z)$, eq. (6.6) reduces to eq. (6.2) and it is an adjustment pair for fig. 6.2(a). Similarly, if the pair is $(Z; ;)$, eq. (6.6) reduces to eq. (6.3) and it is an adjustment pair for fig. 6.2(b). These are two extreme cases where all variables in $\mathbf{Z}$ are measured without bias or where none has been measured externally. Equation (6.6) allows for a more flexible combination. Example 34 (shown later) is a case where a mix of covariates measured with and without selection bias is needed to solve the problem by adjustment.

Building on the concept of *proper causal path* (definition 19) and *proper backdoor graph* (definition 20) introduced in section 6.1, we develop a criterion that determines precisely whether a pair $(\mathbf{Z}; \mathbf{Z}^{\mathbf{T}})$ provides a valid adjustment.

Definition 22 (Generalized Adjustment for Selection Bias)**.** Given a causal diagram G augmented with selection variable S, disjoint sets of variables $\mathbf{X}; \mathbf{Y}; \mathbf{Z}$ and a set $\mathbf{Z}^{\mathbf{T}}$ $\mathbf{Z}$; $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ is an *admissible* pair relative to $\mathbf{X}; \mathbf{Y}$ in G if:

(i) No element in $\mathbf{Z}$ is a descendant in $G_{\overline{\mathbf{X}}}$ of any $W \in \mathbf{X}$ lying on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$.

(ii) All non-causal paths in G from $\mathbf{X}$ to $\mathbf{Y}$ are blocked by $\mathbf{Z}$ and S.

(iii) $\mathbf{Z}^{\mathbf{T}}$ d-separates $\mathbf{Y}$ from S in the proper backdoor graph, i.e. $\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}}$ $_{G^{pbd}_{\mathbf{XY}}}$.

In other words, (i) prevents causal paths to be compromised by conditioning on an element in $\mathbf{Z}$, (ii) requires all non-causal paths to be blocked by $\mathbf{Z}$, and (iii) ensures that the influence of the selection mechanism on the outcome is nullified by $\mathbf{Z}^{\mathbf{T}}$.

The following theorem states that the pairs *admissible* by the graphical criterion in definition 22 are exactly those that constitute *adjustment pairs* as in definition 21.

Theorem 9 (Admissible Pairs are Adjustment Pairs)**.** *$\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ is an* adjustment pair *for $\mathbf{X}; \mathbf{Y}$ in G if and only if it satisfy the generalized adjustment criterion (definition 22).*

The following are some examples of the use of the criterion.

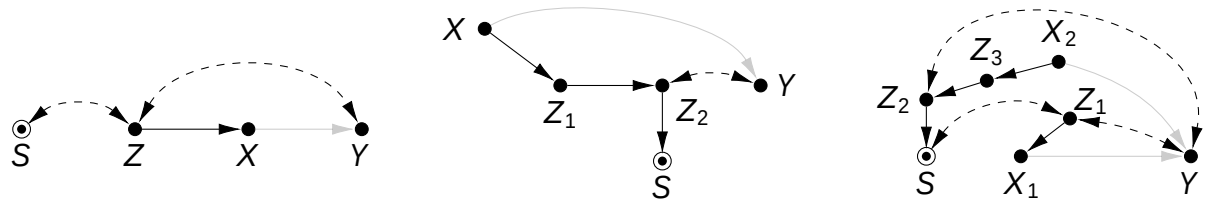
Example 33 (Using the generalized adjustment criterion)**.** Consider the causal diagram in fig. 6.3(a), $(fZg; ;)$ is an adjustment pair for the effect of X on Y. First, Z is not in any causal path and it is not a descendant of any node in such a path. Second, the only noncausal path $X \quad Z$ L99K Y is blocked by Z. Third, Y is d-separated from S given Z. As the three conditions in definition 22, it follows that

$$P(y \mid do(x)) = \sum_{z} P(y \mid x; z; S = 1)P(z \mid S = 1): \tag{6.7}$$

Next, for fig. 6.3(b), $(fZ_1; Z_2g; fZ_2g)$ is an adjustment pair. Both Z_1 and Z_2 are not in any proper causal path, $fZ_1; Z_2g$ block every path in the proper back-door graph, and fZ_2g d-separates S and Y in G. Then,

$$P(y \mid do(x)) = \sum_{z_1;z_2} P(y \mid x; z_1; z_2; S = 1)P(z_1 \mid z_2; S = 1)P(z_2): \tag{6.8}$$

Having unbiased data on all candidate covariates is not necessarily helpful. It may be the case that only valid adjustments require the use of biased measures for a covariate, as in the following example.



(a) (b) (c)

Figure 6.3: Some causal diagrams where the effect $P(y \mid do(x))$ can be recovered by adjustment. The corresponding proper back-door graph can be obtained by removing the edges in gray.

Example 34 (External measurements are not necessarily better)**.** Consider the model in fig. 6.3(c). There are only two admissible adjustment pairs: $(\{Z_1, Z_2, Z_3\}, \{Z_2, Z_3\})$ or $(\{Z_1, Z_2, Z_3\}, \{Z_2\})$, but not one that includes Z_1 in $\mathbf{Z}^{\mathbf{T}}$. Using the latter pair, the query can be identified as

$$P(y \mid do(x)) = \sum_{z_1, z_2, z_3} P(y \mid x, z_1, z_2, z_3) P(z_1, z_3 \mid z_2, S = 1) P(z_2). \tag{6.9}$$

Overall, theorem 9 completely solves the *sbt-ID* task when there exists an adjustment functional mapping the input to the query. In other words, there is no adjustment equal to the causal effect of X on Y or this is recoverable if and only if an admissible pair $\mathbf{Z}, \mathbf{Z}^{\mathbf{T}}$ $\mathbf{T}$ exists.

6.2.1 Listing Admissible Pairs Efficiently

Once the admissibility of adjustment pairs has been characterized, it is natural to ask how to find them systematically and efficiently. This task is especially relevant since factors such as feasibility, cost, and statistical power may be relevant when choosing one of such sets.

To illustrate the complexity of this task, suppose we want to list all possible adjustment

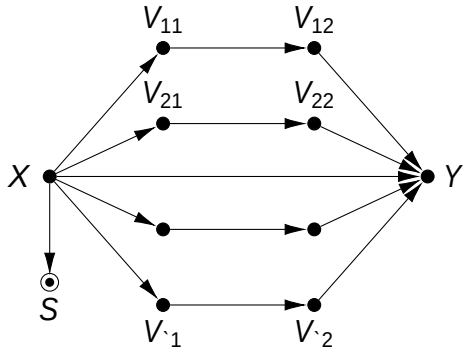


Figure 6.4: Simple diagram where the number of different separators is exponential in the size of the graph

sets for the causal diagram given in fig. 6.4. It contains ℓ non-causal paths from X to Y. For any pair $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ admissible in this model, $\mathbf{Z}$ and $\mathbf{Z}^{\mathbf{T}}$ must contain at least one variable in every one of the ℓ paths. For a path i, either V_{i1}, V_{i2}, or both should be in those sets. In total, there are 3^{ℓ} different $\mathbf{Z}$, and for each one of them there are 3^k sets $\mathbf{Z}^{\mathbf{T}}$, where k is the number of paths that contain both variables in $\mathbf{Z}$. The possible admissible pairs are in the Cartesian product of those sets, which amounts to $O(3^{2\ell})$ possibilities. Any algorithm that aims to output all admissible sets will take exponential time. Hence, no efficient algorithm exists for this task. To ameliorate this problem, we consider a special complexity class called polynomial delay [91]. Algorithms belonging to this class have the special property that the time required to output the first solution (or indicate failure), and the time between the outputs of consecutive solutions, is polynomial in the size of the input.

The following is an alternative and equivalent definition of the criterion given in definition 22. It will prove useful in finding a solution in the polynomial delay class.

Definition 23 (Generalized Adjustment — Procedural Criterion)**.** Given a causal diagram G augmented with selection variable S, disjoint sets of variables $\mathbf{X}; \mathbf{Y}; \mathbf{Z}$ and a set $\mathbf{Z}^{\mathbf{T}}$ $\mathbf{Z}$; $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ is an *admissible* pair relative to $\mathbf{X}; \mathbf{Y}$ in G if:

(a) $\mathbf{Z} \setminus Dpcp(\mathbf{X}; \mathbf{Y}) = ;$

(b) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}; S)_{G^{pbd}_{\mathbf{XY}}}$

(c) $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}})_{G^{pbd}_{\mathbf{XY}}}$

where $Dpcp(\mathbf{X};\mathbf{Y}) = De\ (De(\mathbf{X})_{G_{\overline{\mathbf{X}}}} \cap \mathbf{X}) \setminus An(\mathbf{Y})_{G_{\overline{\mathbf{X}}}}$.

The set $Dpcp(\mathbf{X};\mathbf{Y})$ was originally introduced in [90] to account for the set of descendants of variables that lie in a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$.

Proposition 5. *Definition 23 is equivalent to definition 22.*

In fact, definition 23 is appealing to the task since each of the conditions can be easily verified algorithmically in a graph.

The following definition will be used to describe a collection of sets that separate variables in a causal model, subject to subset and superset constraints:

Definition 24 (Family of Separators)**.** Let $\mathbf{X};\mathbf{Y};\mathbf{R}$ be disjoint sets of variables in a causal diagram G, and let $\mathbf{I} \quad \mathbf{R}$ be another set. Define

$$\mathcal{Z}_G(\mathbf{X};\mathbf{Y})\langle \mathbf{I};\mathbf{R}\rangle := \{\mathbf{Z} \mid (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G \text{ and } \mathbf{I} \quad \mathbf{Z} \quad \mathbf{R}\} \tag{6.10}$$

to be the family of all sets $\mathbf{Z}$ that d-separate $\mathbf{X}$ and $\mathbf{Y}$ in G and contain all elements in $\mathbf{I}$ but no element outside $\mathbf{R}$.

For convenience, let the set of viable candidates for adjustment be denoted and defined as:

$$\mathcal{C} = \mathbf{V} \cap (\mathbf{X} \cup \mathbf{Y} \cup Dpcp(\mathbf{X};\mathbf{Y})) \tag{6.11}$$

Using this notation, the families that satisfy the conditions of our criterion can be specified. For conditions (a) and (b):

$$\mathcal{Z}_{a;b} = \left\{ \mathbf{Z} \;\middle|\; \mathbf{Z} \cup \{Sg \in \mathcal{Z}_{G^{pbd}_{\mathbf{XY}}}(\mathbf{X};\mathbf{Y})\langle Sg; \mathcal{C} \cup \{Sg\}\rangle \right\} \tag{6.12}$$

The algorithm should take into account the availability of external data over a set of covariates $\mathbf{T}$. To obtain admissible pairs for which the adjustment is estimable, the set $\mathbf{T}$ is

incorporated in the definition of the family for item (c):

$$\mathcal{Z}_c = \mathcal{Z}_{G^{pbd}_{\mathbf{XY}}}(fSg; \mathbf{Y})\, h;; \mathbf{T}i : \tag{6.13}$$

Our task can be summarized as finding pairs in the set:

$$\mathcal{Z}_{a;b;c} = \quad \mathbf{Z}; \mathbf{Z}^{\mathbf{T}} \quad 2\ \mathcal{Z}_{a;b} \quad \mathcal{Z}_c \quad \mathbf{Z}^{\mathbf{T}} \quad \mathbf{Z} \tag{6.14}$$

Algorithm 9 presents the procedure LISTADJPAIRS that solves this problem, as well as the auxiliary routines used by it. Specifically, it may be used to:

1. Given external data $P(\mathbf{t})$, list all admissible pairs such that $\mathbf{Z}^{\mathbf{T}} \quad \mathbf{T}$.
2. List all admissible pairs (by setting $\mathbf{T} = \mathbf{V}\ n\ (\mathbf{X}\ [\ \mathbf{Y})$) such that scientists know what external data to measure.

Functions LISTSEPAB and LISTSEPC are modifications of the enumeration algorithm LISTSEP in [90]. The function FINDSEP is also described in that paper, and works as follows: given a causal diagram G, sets of variables $\mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{R}$, where $\mathbf{X}; \mathbf{Y}; \mathbf{R}$ are disjoint and $\mathbf{I} \quad \mathbf{R}$; FINDSEP is guaranteed to output a $\mathbf{Z} \in 2\ \mathcal{Z}_G(\mathbf{X}; \mathbf{Y})\, h\mathbf{I}; \mathbf{R}i$ whenever there exists a separator $\mathbf{C}$ such that $\mathbf{I} \quad \mathbf{C} \quad \mathbf{R}$; otherwise it returns $?$ denoting failure.

Proposition 6 (Correctness of LISTSEPC)**.** *Given a graph G, a variable S, sets of variables $\mathbf{Y}; \mathbf{I}; \mathbf{R}; \mathbf{Z}$, where $fSg; \mathbf{Y}; \mathbf{Z}$ are disjoint and $\mathbf{I} \quad \mathbf{R} \quad \mathbf{Z}$;* LISTSEPC *outputs all pairs $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$, where $\mathbf{Z}^{\mathbf{T}} \in 2\ \mathcal{Z}_G(fSg; \mathbf{Y})\, h\mathbf{I}; \mathbf{R}i$.*

Proposition 7 (Correctness of LISTSEPAB)**.** *Given a graph G, a variable S, sets of variables $\mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{R}; \mathbf{T}$, where $\mathbf{X}; \mathbf{Y}; fSg; \mathbf{R}$ are disjoint, $\mathbf{I} \quad \mathbf{R}$ and $\mathbf{T} \quad \mathcal{C}$;* LISTSEPAB *outputs all pairs in $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}} \in 2\ \mathcal{Z}_G(\mathbf{X}; \mathbf{Y})\, h\mathbf{I}; \mathbf{R}i \quad \mathcal{Z}_G(fSg; \mathbf{Y})\, h;; \mathbf{T}i \quad \mathbf{Z}^{\mathbf{T}} \quad \mathbf{Z}$.*

The following theorem states that the algorithm LISTADJPAIRS can solve the task proposed in this section.

Algorithm 9 LISTADJPAIRS($G; \mathbf{X}; \mathbf{Y}; S; \mathbf{V}; \mathbf{T}$)

1: $G^{pbd}_{\mathbf{XY}}$ Compute proper backdoor graph from G
2: $\mathbf{R}$ $(\mathbf{V} \cup \{S\}) \setminus (\mathbf{X} \cup \mathbf{Y} \cup Dpcp(\mathbf{X}; \mathbf{Y}))$
3: LISTSEPAB($G^{pbd}_{\mathbf{XY}}; \mathbf{X}; \mathbf{Y}; S; \{S\}; \mathbf{R}; \mathbf{T}$)

Theorem 10 (Correctness of LISTADJPAIRS)**.** *Given a graph G, disjoint sets $\mathbf{X}; \mathbf{Y}; \mathbf{T}$, and a selection variable S,* LISTADJPAIRS *outputs all admissible pairs $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ relative to $\mathbf{X}; \mathbf{Y}$ in G such that $\mathbf{Z}^{\mathbf{T}} \not\subseteq \mathbf{T}$.*

It is worth noting that a straightforward adaptation of the algorithm LISTSEP [90] may be used to find sets in $\mathbf{Z} \in \mathcal{Z}_{a;b}$ and $\mathbf{Z}^{\mathbf{T}} \in \mathcal{Z}_c$. However, the condition $\mathbf{Z}^{\mathbf{T}} \not\subseteq \mathbf{Z}$ has to be verified to produce admissible pairs. One strategy could be to search for sets in $\mathcal{Z}_{a;b}$ first, and then, while a second run outputs each set in $\mathcal{Z}_c$, validate if it is a subset of any output from the first batch of sets. In the worst case, exponential time is required to output the first admissible pair. A better idea would be to search for sets in $\{\mathbf{Z}^{\mathbf{T}} \in \mathcal{Z}_c \mid \mathbf{Z}^{\mathbf{T}} \not\subseteq \mathbf{Z}\}$ as soon as some $\mathbf{Z} \in \mathcal{Z}_{a;b}$ is found, and then output pairs made of $\mathbf{Z}$ and the outputs of the secondary search. While improving over the original strategy, it may be the case that, for some sets in $\mathcal{Z}_{a;b}$, there is no set in $\mathcal{Z}_c$, which would lead to an exponential waiting time to get the first output.

Proposition 8 and theorem 11 show that LISTADJPAIRS is, in fact, able to achieve $O(n(n+m))$ delay by carefully combining the search for the components of the pairs, where $n; m$ are the number of variables and edges in G, respectively.

Proposition 8 (Complexity of LISTSEPAB)**.** LISTSEPAB *works with $O(n(n+m))$ delay.*

Theorem 11 (Complexity of LISTADJPAIRS)**.** LISTADJPAIRS *outputs all admissible pairs such that $\mathbf{Z}^{\mathbf{T}} \not\subseteq \mathbf{T}$ with $O(n(n+m))$ polynomial delay.*

Using covariates from $An(\mathbf{X} \cup \mathbf{Y})$ is sufficient to block any biasing path when controlling for confounding bias, which does not hold when selection bias comes into play. The proposition below constitutes a natural extension of this result when searching for adjusting pairs, in particular, considering the set $An(\mathbf{X} \cup \mathbf{Y} \cup \{S\})$.

Algorithm 10 LISTSEPAB($G; \mathbf{X}; \mathbf{Y}; S; \mathbf{I}; \mathbf{R}; \mathbf{T}$)

if FINDSEP$G; fSg; \mathbf{Y}; \mathbf{I}; \mathbf{R} \not\perp ?$ ^ FINDSEP$G; fSg; \mathbf{Y}; ;; \mathbf{R} \setminus \mathbf{T} \not\perp ?$ **then**
  if $\mathbf{I} = \mathbf{R}$ **then**
    LISTSEPC($G; S; \mathbf{Y}; ;; \mathbf{I} \setminus \mathbf{T}; \mathbf{I} \cap fSg$)
  else
    V arbitrary variable from $\mathbf{R} \cap \mathbf{I}$
    LISTSEPAB($G; \mathbf{X}; \mathbf{Y}; \mathbf{I} \cup fVg; \mathbf{R}; \mathbf{T}$)
    LISTSEPAB($G; \mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{R} \cap fVg; \mathbf{T}$)
  end if
end if

Algorithm 11 LISTSEPC($G; \mathbf{X}; \mathbf{Y}; S; \mathbf{I}; \mathbf{R}; \mathbf{T}$)

if FINDSEP$G; fSg; \mathbf{Y}; \mathbf{I}; \mathbf{R} \not\perp ?$ **then**
  if $\mathbf{I} = \mathbf{R}$ **then**
    output $(\mathbf{Z}; \mathbf{I})$
  else
    V arbitrary variable from $\mathbf{R} \cap \mathbf{I}$
    LISTSEPC($G; \mathbf{X}; \mathbf{Y}; \mathbf{I} \cup fVg; \mathbf{R}; \mathbf{Z}$)
    LISTSEPC($G; \mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{R} \cap fVg; \mathbf{Z}$)
  end if
end if

Proposition 9. *Suppose a pair* $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ *is admissible relative to* $\mathbf{X}; \mathbf{Y}$ *in* G. *Then, the pair* $\mathbf{Z}_{\mathbf{A}}; \mathbf{Z}^{\mathbf{T}}_{\mathbf{A}}$, *where* $\mathbf{Z}^{\mathbf{T}}_{\mathbf{A}} = \mathbf{Z}^{\mathbf{T}} \setminus An(\mathbf{X} \cup \mathbf{Y} \cup fSg)$ *and* $\mathbf{Z}_{\mathbf{A}} = \mathbf{Z} \setminus An(\mathbf{X} \cup \mathbf{Y} \cup fSg)$, *is also admissible.*

If the data scientist is not interested in deciding among different adjustment pairs to use, it is possible to explicitly construct a pair if one exists, namely:

Theorem 12 (Explicit admissible set construction)**.** *There exists an admissible pair in a causal diagram* G *relative to disjoint sets of variables* $\mathbf{X}; \mathbf{Y}$ *if and only if the pair* $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ *is admissible, where*

$$\mathbf{Z} = An(\mathbf{X} \cup \mathbf{Y} \cup fSg)_{G^{pbd}_{\mathbf{XY}}} \setminus \mathcal{C} \tag{6.15}$$

$$\mathbf{Z}^{\mathbf{T}} = (An(fSg \cup \mathbf{Y})_{G^{pbd}_{\mathbf{XY}}} \setminus \mathbf{T}) \setminus \mathcal{C} \tag{6.16}$$

Moreover, theorem 12 also allows one to determine the existence of an admissible pair and construct one in $O(n + m)$ time.

6.3 Covariate Adjustment for Generalizing Experimental Findings[2]

The previous section described an adjustment criterion for recovering causal effects from a combination of selection biased data and unbiased measurements. There are, however, other generalization tasks that involve selection bias and can also be solved by adjustment. In this section, we look at the problem of generalizing an interventional distribution obtained by controlled experimentation. That is, instead of trying to assess a causal effect from selection-biased data, the challenge is to generalize a causal effect from a source domain to a target domain, while the sampling in the controlled experiment is not random.

A properly carried-out experiment will effectively control for confounding bias, and the effect of the treatment $\mathbf{X}$ on the outcome $\mathbf{Y}$ will be valid for *the population represented in the experiment*, say, domain . However, the goal is usually not to make statements about the units involved in the experiment, but to generalize the findings to a (usually much) larger and possibly different population (domain). Invalid conclusions about the target population will be reached if the generalization biases are left uncontrolled. In other words, $P(\mathbf{y} \mid do(\mathbf{x}))$, obtained in may differ significantly from $P(\mathbf{y} \mid do(\mathbf{x}))$, the corresponding causal quantity for the target population .

Example 35 (Using covariate adjustment to generalize a causal effect)**.** The selection diagram in fig. 6.5 represents the situation described in example 1 (chapter 1), where the FDA was trying to assess the effect of the use of antidepressants on the risk of suicide attempts. Suppose the aim was to obtain the effect $P(y \mid do(x))$ in domain (general population) from the data $P(y, b, e \mid do(x), S=1)$ coming from the domain (controlled group). In practice, experimental data from the source domain may be insufficient to

[2]This section is based on the paper [38]

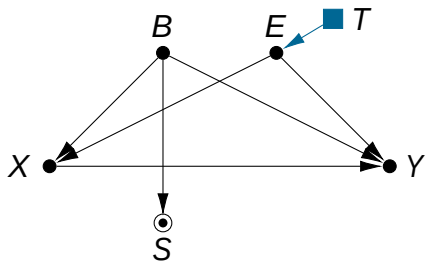


Figure 6.5: Selection diagrams with $\mathbf{T}$ and S nodes indicating differences between populations and the sampling selection mechanism.

identify the target effect. Still, it is not uncommon that nonexperimental, unbiased data may be available in the target population, at least over some subset of the variables, $\mathbf{W} \subseteq \mathbf{V}$ (i.e., $P(\mathbf{W})$) (e.g., data coming from the census). In these situations, covariate adjustment provides a natural way of combining data from the two domains. For example, if $P(B, E)$ is available in the target population, then the target effect can be obtained as

$$P(y \mid do(x)) = \sum_{b,e} P(y \mid x, b, e, S = 1) P(b, e), \tag{6.17}$$

which is a function of both the data obtained in the randomized experiment and the data observed in the overall population.

The signature of the task considered in this section is

$$I_{exp\text{-}TR} = \langle P(\mathbf{y} \mid do(\mathbf{x})), \mathbf{P}, G^{\Delta} \rangle, \tag{6.18}$$

where $\mathbf{P} = \{P(\mathbf{V} \mid do(\mathbf{x}), S = 1), P(\mathbf{W})\}$. In other words, given qualitative causal assumptions in the form of a selection diagram G^{Δ}, data $P(\mathbf{V} \mid do(\mathbf{x}), S = 1)$ in domain Π, and $P(\mathbf{V})$ in domain Π^*, we would like to determine if $P(\mathbf{y} \mid do(\mathbf{x}))$ is estimable by adjustment on a set $\mathbf{Z} \subseteq \mathbf{W} \subseteq \mathbf{V}$. We are looking for sufficient and necessary conditions to determine if the adjustment-like expression

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, S = 1) P(\mathbf{z}), \tag{6.19}$$

holds based on the assumptions encoded in a selection diagram G^{Δ}. The right hand side of Eq. (6.19) contains two terms corresponding to different distributions – the first is the experimental one from the source ( ) that may be affected by selection bias; the second is the distribution over a set of covariates measured in the target domain ( ).

For convenience, when considering a set $\mathbf{Z}$ and treatment $\mathbf{X}$, let $\mathbf{Z}_{\mathrm{nd}} = \mathbf{Z} \cap De(\mathbf{X})$ denote the non-descendants of $\mathbf{X}$ in $\mathbf{Z}$, and $\mathbf{Z}_{\mathrm{d}} = \mathbf{Z} \setminus De(\mathbf{X})$ denote the descendants of $\mathbf{X}$. It turns out that conditioning on variables from $\mathbf{Z}_{\mathrm{d}}$ that are independent of the outcome $\mathbf{Y}$ given $\mathbf{Z}_{\mathrm{nd}}$ in the experimental distribution does not introduce spurious correlation into the adjustment. On the other hand, we need to pay special attention to those variables in $\mathbf{Z}_{\mathrm{d}}$ d-connected with $\mathbf{Y}$ in the interventional graph $G_{\overline{\mathbf{X}}}$ (given $\mathbf{X}$), that we will denote as

$$\mathbf{Z}_{\mathrm{p}} = \left\{ Z \in \mathbf{Z}_{\mathrm{d}} \;\middle|\; (Z \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}_{\mathrm{nd}}, \mathbf{X})_{G_{\overline{\mathbf{X}}}} \right\}. \tag{6.20}$$

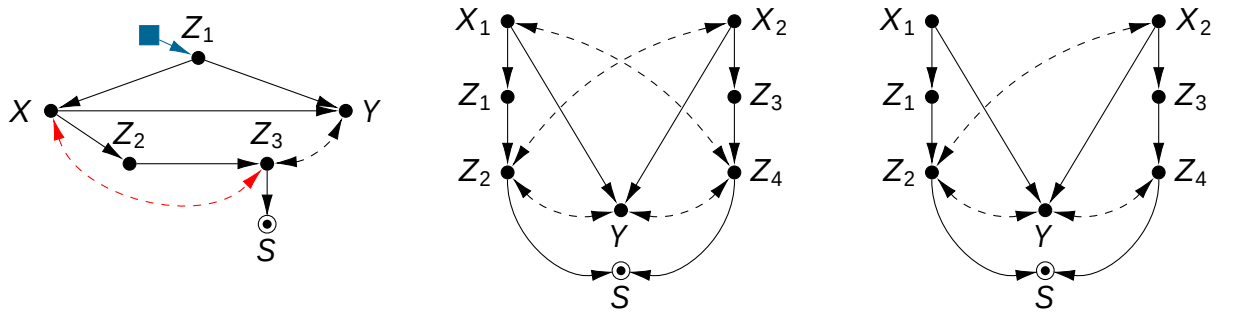
We devised a graphical condition [38] to characterize the sets $\mathbf{Z}$ that yield valid adjustments for $P(\mathbf{y} \mid do(x))$, where X is a single variable.

Definition 25 (Generalization Adjustment (st-adjustment) Criterion (singleton treatment))**.** Given a selection diagram G^{Δ} with transportability and selection bias variables, respectively, $\mathbf{T}$ and S, relative to domains and , a treatment X, and disjoint sets $\mathbf{Y}, \mathbf{Z}$ $\mathbf{V}$, the set $\mathbf{Z}$ is said to satisfy the st-adjustment criterion relative to $(X, \mathbf{Y})$ in G^{Δ} if

(i) The variables in $\mathbf{Z}_{\mathrm{p}}$ are independent of the treatment given all other covariates, i.e., $(\mathbf{Z}_{\mathrm{p}} \not\perp\!\!\!\perp X \mid \mathbf{Z} \cap \mathbf{Z}_{\mathrm{p}})$.

(ii) The outcome is independent of the transportability nodes and the selection bias mechanism given the covariates and X, i.e., $(\mathbf{Y} \not\perp\!\!\!\perp \mathbf{T}, S \mid \mathbf{Z}, X)_{G^{\Delta}_{\overline{X}}}$.

Since the variables in $\mathbf{Z}_{\mathrm{p}}$ are correlated with the outcome (by definition), the first condition requires them to be independent of the treatment X, given the other covariates, to prevent spurious correlation or the disturbance of causal paths when employing such variables. The second condition accounts for the generalizability issues — it requires



(a) Case where no set satisfies st-adjustment.

(b) No order over Z_1, Z_2, Z_3, Z_4 is suitable for adjustment.

(c) Order $Z_1 < Z_2 < Z_3 < Z_4$ is suitable for adjustment.

Figure 6.6: Models with multiple treatment variables $\mathbf{X} = \{X_1, X_2\}$.

the outcome to be independent of the transportability (T) and selection bias nodes (S) in the effect specific to the levels of the set $\mathbf{Z}$; the criterion owes its name, st-adjustment, to this condition. In contrast to similar criteria, no condition is required for controlling confounding due to the experimental nature of the data. For fig. 6.5 let $\mathbf{Z} = \{B, E\}$, then $\mathbf{Z}_{\mathbf{p}} = \emptyset$; since neither B nor E are descendants of X, and the first condition is satisfied. For the second conditions, one can immediately verify that $(Y \perp\!\!\!\perp T, S \mid B, E, X)_{G_{\overline{X}}}$ holds.

Consider the diagram in fig. 6.6(a). For the set $\mathbf{Z} = \{Z_1\}$, condition (i) is trivially satisfied because $\mathbf{Z}_{\mathbf{p}} = \emptyset$. However, there is an active path $S \leftarrow Z_3 \dashleftarrow Y$ that violates (ii). In fact, Z_3 needs to be included in $\mathbf{Z}$, but then $(Z_3 \not\perp\!\!\!\perp X \mid Z_1)$ because of the directed path $X \rightarrow Z_2 \rightarrow Z_3$. We have to include Z_2 in $\mathbf{Z}$ to block this path, which leads to the same $\mathbf{Z}_{\mathbf{p}}$, but now there is still a path $X \dashleftarrow Z_3$ that violates the first condition. It turns out, there is no set $\mathbf{Z}$ satisfying the criterion for this case. The st-adjustment criterion is not only sufficient but also necessary.

Theorem 13 (st-adjustment (singleton treatment))**.** *Given a selection diagram G^{Δ}, a singleton X, and disjoint sets $\mathbf{Y}$ and $\mathbf{Z}$, the causal effect $P(\mathbf{y} \mid do(x))$ is given by*

$$P(\mathbf{y} \mid do(x)) = \sum_{\mathbf{z}} P(\mathbf{y} \mid do(x), \mathbf{z}, S=1) P(\mathbf{z}) \tag{6.21}$$

if and only if $\mathbf{Z}$ satisfies the st-adjustment criterion relative to $(X, \mathbf{Y})$.

Even though controlling for one treatment variable at a time may be sufficient in some applications, in practice, there are settings where multiple factors need to be tested concurrently. In this section, we address more challenging settings involving causal effects of multiple treatment variables. For example, in online marketing, experiments are used to test the effectiveness of a combination of variables such as content position, media, and audience, on user interaction, clicks, or conversion. Due to the cost and user participation required to carry out these experiments, it is desirable to be able to generalize them to alternative audiences and correct for sampling issues.

To handle multiple treatments, adjusting for the descendants of $\mathbf{X}$ may again induce a spurious correlation between $\mathbf{X}$ and $\mathbf{Y}$. More attention is needed to the variables in $\mathbf{Z_p}$ (defined in eq. (6.20)) and how they are related to the multiple treatments $\mathbf{X}$.

Example 36 (Subtleties with adjusting for multiple treatments)**.** Consider the models in fig. 6.6(b) and fig. 6.6(b) and set $\mathbf{Z} = \{Z_1, Z_2, Z_3, Z_4\}$[3] leading to $\mathbf{Z_p} = \{Z_2, Z_4\}$. Note that $\mathbf{Z_p}$ is not independent of $\mathbf{X} = \{X_1, X_2\}$ given $\mathbf{Z} \setminus \mathbf{Z_p} = \{Z_1, Z_3\}$ in either one of the diagrams, hence condition (i) of definition 25 fails in both cases. Even so, there is a subtle difference between the two models: while adjusting for $\mathbf{Z}$ is not valid in fig. 6.6(b), it is

[3]The two selection diagrams do not have $\mathbf{T}$ nodes, meaning the populations are the same in source and target domains with only selection bias issue occurring.

guaranteed to yield $P(y \mid do(\mathbf{x}))$ in fig. 6.6(c). To witness, see the following derivation

$$
\begin{aligned}
P(y \mid do(\mathbf{x})) &= P(y \mid do(\mathbf{x})) \sum_{z_1} P(z_1) && (6.22)\\
&= \sum_{z_1} P(y \mid do(\mathbf{x}); z_1) P(z_1) && (6.23)\\
&= \sum_{z_1, z_2} P(y \mid do(\mathbf{x}); z_1, z_2) P(z_2 \mid do(\mathbf{x}); z_1) P(z_1) && (6.24)\\
&= \sum_{z_1, z_2} P(y \mid do(\mathbf{x}); z_1, z_2) P(z_1, z_2) && (6.25)\\
&= \sum_{z_1, z_2, z_3} P(y \mid do(\mathbf{x}); z_1, z_2, z_3) P(z_1, z_2, z_3) && (6.26)\\
&= \sum_{\mathbf{z}} P(y \mid do(\mathbf{x}); \mathbf{z}) P(z_4 \mid do(\mathbf{x}); z_1, z_2, z_3) P(z_1, z_2, z_3) && (6.27)\\
&= \sum_{\mathbf{z}} P(y \mid do(\mathbf{x}); \mathbf{z}) P(\mathbf{z}) && (6.28)\\
&= \sum_{\mathbf{z}} P(y \mid do(\mathbf{x}); \mathbf{z}, S = 1) P(\mathbf{z}) && (6.29)
\end{aligned}
$$

We first introduce Z_1 into the adjustment (eq. (6.22)) using the fact that is d-separated of Y given $\mathbf{X}$ in $G^{\Delta}_{\overline{\mathbf{X}}}$, hence it does not introduce any spurious correlation eq. (6.23). Next, we added Z_2 by conditioning eq. (6.24), and since X_2 has no effect on $\{Z_1, Z_2\}$, $P(z_2 \mid do(\mathbf{x}); z_1) = P(z_2 \mid do(x_1); z_1)$. Also, given Z_1, Z_2 is independent of X_1, so no spurious correlation is added eq. (6.25). Similarly, Z_1, Z_3 is independent of Y given the already introduced $\{Z_1, Z_2\}$ eq. (6.26). Finally, Z_4 is independent of $\{X_1, X_2\}$ given $\{Z_1, Z_2, Z_3\}$ eq. (6.28). After both Z_2 and Z_4 have been adjusted for, the outcome is independent of the selection mechanism S, and the causal effect can be expressed in the form of the st-adjustment eq. (6.29).

Remarkably, no other set $\mathbf{Z}$ is valid for adjustment in this model, and the steps described can only be performed in the given order. As a matter of fact, the reason why $\mathbf{Z}$ will not work for fig. 6.6(b) is that in the last step, we have a distribution $P(z_4 \mid do(\mathbf{x}); z_1, z_2, z_3)$ and since X_1 has a causal effect over $\{Z_1, Z_2\}$, this conditional probability is not guaranteed to be equal to $P(z_4 \mid do(x_2); z_1, z_2, z_3)$, if it was, we could employ $(Z_4 \perp\!\!\!\perp X_2 \mid Z_1, Z_2, Z_3)$ to finish the derivation. A symmetric problem with Z_2 arises If we change the order so that

Z_4 is added before $\{Z_1, Z_2\}$.

To solve this general version of the problem, we first consider only the source domain, that is, the conditions for $P^*(y \mid do(x))$ to be computable in the form of eq. (6.28).

Definition 26 (Experimental Adjustment (e-adjustment) Criterion)**.** Given a causal diagram G and disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, $\mathbf{Z}$ is said to satisfy the e-adjustment criterion relative to $(\mathbf{X}, \mathbf{Y})$ in G if there exists an order over $\mathbf{Z}$: $Z_1 < Z_2 < \cdots$, such that $\mathbf{Z}_{nd} < \mathbf{Z}_d$, and for each $Z_i \in \mathbf{Z}_d$ we have

$$Z_i \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}^{i-1}, \mathbf{X} \quad \text{in } G_{\overline{\mathbf{X}}}, \text{ or} \tag{6.30}$$

$$Z_i \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}^{i-1} \quad \text{in } G_{\overline{\mathbf{X}(\mathbf{Z}^{i-1})}}, \tag{6.31}$$

where $\mathbf{Z}^{i}$ denotes the set $\{Z_1, \ldots, Z_i\}$.

Note that although it may seem computationally expensive to determine the existence of an ordering over $\mathbf{Z}$ satisfying e-adjustment, we showed in [38] that can be verified efficiently. Also, if $\mathbf{Z}_p$ is empty, definition 26 is trivially satisfied. The following theorem ties the definition of e-adjustment with the adjustment expression.

Theorem 14. *Given a causal diagram G and disjoint sets of variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, the distribution $P(y \mid do(x))$ is given by*

$$P(y \mid do(x)) = \sum_{z} P(y \mid do(x), z) P(z) \tag{6.32}$$

if and only if $\mathbf{Z}$ satisfies the e-adjustment criterion relative to $(\mathbf{X}, \mathbf{Y})$.

Leveraging e-adjustment, we characterized adjustment sets allowing for the generalization of experiments across domains with multiple treatments.

Definition 27 (st-adjustment criterion (multiple treatments))**.** Given a selection diagram G^{Δ} with transportability and selection bias variables, respectively, $\mathbf{T}$ and S, relative

to domains and , and disjoint sets $\mathbf{X}; \mathbf{Y}; \mathbf{Z}$ $\mathbf{V}$, the set $\mathbf{Z}$ is said to satisfy the st-adjustment criterion relative to $(\mathbf{X}; \mathbf{Y})$ in G^{Δ} if

(i) $\mathbf{Z}$ satisfies the e-adjustment criterion (definition 26), and

(ii) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{T}; S \mid \mathbf{Z}; \mathbf{X})_{G^{\Delta}_{\overline{\mathbf{X}}}}$.

Theorem 15 (st-adjustment)**.** *Given a selection diagram G^{Δ} and disjoint sets of variables $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, the causal effect $P(\mathbf{y} \mid do(\mathbf{x}))$ is given by*

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}; S=1) P(\mathbf{z}) \tag{6.33}$$

if and only if $\mathbf{Z}$ satisfies the st-adjustment criterion relative to $(\mathbf{X}; \mathbf{Y})$.

In fig. 6.6(c), the set $\mathbf{Z} = \{Z_1; Z_2; Z_3; Z_4\}$ satisfies the st-adjustment criterion with order $Z_1 < Z_2 < Z_3 < Z_4$. Therefore $P(y \mid do(x_1; x_2))$ can be computed by eq. (6.33) as explicitly derived in eq. (6.22)–eq. (6.29).

Theorem 15 implies that the st-adjustment criterion is a complete characterization of valid adjustment sets.

6.3.1 Verifying e-adjustment Efficiently

Evaluating condition (i) of the st-adjustment (definition 27), that is, the existence of an ordering over $\mathbf{Z}$ satisfying e-adjustment (definition 26), may seem computationally hard. However, we will show in this section that it can be verified efficiently by first establishing some properties of e-adjustment.

Lemma 6. *A set $\mathbf{Z}$ satisfies e-adjustment if and only if there exists $Z_i \in \mathbf{Z}$ such that Z_i satisfies eq.* (6.30) *or eq.* (6.31)*, and $\mathbf{Z} \setminus \{Z_i\}$ satisfies e-adjustment.*

Lemma 6 provides a recursive characterization of the order condition. Based on this result, we can verify the existence of an order by finding, at each step, any variable satisfying eq. (6.30) or eq. (6.31) in the set and removing it, as described next:

Algorithm 12 IsEAdmissible($G; \mathbf{X}; \mathbf{Y}; \mathbf{Z}$)

Require: causal diagram G, disjoint subsets $\mathbf{X}; \mathbf{Y}; \mathbf{Z} \quad \mathbf{V}$.
Ensure: **true** if $\mathbf{Z}$ satisfies e-adjustment, **false** otherwise.
1: **if** $\mathbf{Z} \setminus De(\mathbf{X}) = \ ;$ **then**
2: **return true**
3: **end if**
4: **for each** $Z \ 2 \ \mathbf{Z} \setminus De(\mathbf{X})$ **do**
5: **if** $(Z \ ?? \ \mathbf{Y} \ j \ \mathbf{Z} \ n \ fZg; \mathbf{X})_{G_{\overline{\mathbf{X}}}}$ **or**
$(Z \ ?? \ \mathbf{X} \ j \ \mathbf{Z} \ n \ fZg)_{G_{\overline{\mathbf{X}(\mathbf{Z} n fZg)}}}$ **then**
6: **return** *IsEAdmissible*$(G; \mathbf{X}; \mathbf{Y}; \mathbf{Z} \ n \ fZg)$
7: **end if**
8: **end for**
9: **return false**

Lemma 7. *If $\mathbf{Z}$ satisfies e-adjustment, then for any $Z_i \ 2 \ \mathbf{Z}$ satisfying eq.* (6.30) *or eq.* (6.31), *the set $\mathbf{Z} n fZ_i g$ satisfies e-adjustment.*

Leveraging these results, we introduce an algorithm called *IsEAdmissible* (algorithm 12) that efficiently checks if $\mathbf{Z}$ satisfies the e-adjustment criterion.

Theorem 16. *$\mathbf{Z}$ satisfies e-adjustment (definition 26) w.r.t. $(\mathbf{X}; \mathbf{Y})$ in G if and only if IsEAdmissible (algorithm 12) returns true.*

To illustrate how *IsEAdmissible* works, consider again the diagram in fig. 6.6(c) with the set $fZ_1; Z_2; Z_3; Z_4 g$. Line 5 will evaluate true only for Z_4, then the process reduces to verifying if $fZ_1; Z_2; Z_3 g$ has an order. Next, the same condition will evaluate true for Z_3 reducing the problem to $fZ_1; Z_2 g$. The process continues by removing Z_2 and after removing Z_1 the condition on line 1 is satisfied, so line 2 executes and returns true. In the case of fig. 6.6(b) also with $fZ_1; Z_2; Z_3; Z_4 g$, none of the variables in the set will satisfy the condition in line 5 and line 9 returns false.

Let n and m stand, respectively, for the number of variables and edges in the graph. Then *IsEAdmissible* performs at most $n^2 \quad n$ conditional independence tests. Constructing the graphs $G_{\overline{\mathbf{X}}}$ and $G_{\overline{\mathbf{X}(\mathbf{Z} n fZg)}}$, as well as determining the descendants of $\mathbf{X}$ is achievable in $O(n + m)$ time. Testing an independence in the graph can be done in $O(n + m)$ [90]. Therefore, the overall time complexity of *IsEAdmissible* is $O(n^2(n + m))$.

6.3.2 Enumerating Valid Sets for st-adjustment

Armed with a graphical condition to test if a set $\mathbf{Z}$ is valid for adjustment, the natural question is how to find sets satisfying the st-adjustment criterion systematically, as efficiently as possible. In practice, what variables are suitable for adjustment may be determined by factors such as feasibility, cost, and effort required to measure such variables and the quality and quantity of obtainable samples. In this section, we assume the data is available in the target domain over a set $\mathbf{W}$ of variables and our task here is to list all sets $\mathbf{Z} \quad \mathbf{W}$ satisfying the st-adjustment.

The number of sets satisfying the st-adjustment is possibly exponential depending on the topology of the diagram and the target effect. In this sense, it is impossible to construct a procedure that runs in polynomial time since just outputting an exponential number of answers takes exponential time.

Under these conditions, the best guarantee we can provide is that the time to output the first valid set or indicate failure (if there is no satisfying set), and the time between consecutive outputs, is polynomial. Algorithms with this property are said to run with *polynomial delay* [91].

We have developed the algorithm *ListGAdjSets* (algorithm 13) which systematically lists valid adjustment sets, using the recursive subroutine *ListGAdjIR*. *ListGAdjIR* outputs all sets $\mathbf{Z}$, $\mathbf{I} \quad \mathbf{Z} \quad \mathbf{R}$, that satisfy the st-adjustment. At each step, it chooses a variable A and splits the problem into two: listing sets containing A (line 10) and those with no A (line 11), while pruning branches that yield no valid sets (lines 4,13). This strategy is similar to those used in [91], [90], and [36] for listing separating sets in a graph. Here, it has been augmented to recognize the conditions in definition 27 (See appendix. B for details).

Theorem 17. *ListGAdjSets on input* $G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{W}$, *lists all sets* $\mathbf{Z} \quad \mathbf{W}$ *satisfying st-adjustment relative to* $\mathbf{X}; \mathbf{Y}$ *in* G^{Δ}, *with* $O(n^4(n+m))$ *delay.*

Algorithm 13 ListGAdjSets($G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{W}$)

Require: selection diagram G^{Δ} over variables $\mathbf{V}$ and indicators $\mathbf{T}$, S; disjoint subsets of $\mathbf{X}; \mathbf{Y}; \mathbf{W} \subseteq \mathbf{V}$.

Ensure: list of subsets $\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_k \subseteq \mathbf{W}$ satisfying definition 27

$\mathbf{F} \leftarrow De(\mathbf{X})_{G^{\Delta}_{\overline{\mathbf{X}}}} \cap \mathbf{X} \setminus An(\mathbf{Y})_{G^{\Delta}_{\overline{\mathbf{X}}}}$
$\mathbf{R} \leftarrow \mathbf{W} \cap (\mathbf{X} \cup \mathbf{Y} \cup \mathbf{F})$
ListGAdjIR $G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{T}; S; \emptyset; \mathbf{R}$

function ListGAdjIR($G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{R}$)
  if *ExistsSep* $G^{\Delta}_{\overline{\mathbf{X}}}; \mathbf{T} \cup \{S\}; \mathbf{Y}; \mathbf{I}; \mathbf{R}$ **then**
    if $\mathbf{I} = \mathbf{R}$ **then**
      print $\mathbf{I}$
    else
      $A \leftarrow$ variable from $(\mathbf{R} \cap \mathbf{I})$ such that
        $A \in De(\mathbf{X})$, else
        $(A \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{I}; \mathbf{X})_{G^{\Delta}_{\overline{\mathbf{X}}}}$, else
        IsEAdmissible$(G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{I} \cup \{A\})$
      if A exists **then**
        ListGAdjIR $G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{I} \cup \{A\}; \mathbf{R}$
        ListGAdjIR $G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{R} \cap \{A\}$
      else
        ListGAdjIR $G^{\Delta}; \mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{I}$
      end if
    end if
  end if

6.4 Systematic Recoverability from Selection Biased Data[4]

In this section, we extend the algorithmic framework introduced in chapter 3 to support tasks including selection biased distributions as inputs, and goes beyond adjustment. First, we start from a task that only considers the availability of selection biased data [27] with the following signature:

$$I_{sb\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})); \{P(\mathbf{V} \mid S = 1)\}; \mathcal{G} \rangle. \tag{6.34}$$

Solving the *sb-ID* task requires the ability to derive c-factors from a distribution of the form $P(\mathbf{V} \mid S = 1)$. The following operators are defined for this purpose.

[4]This section is based on the paper [37].

Lemma 8 (-, ^{0}-, and -operator for selection-biased c-factors)**.** *Let* $\mathbf{T}$ $\mathbf{V}$, $\mathbf{C}$ $\mathbf{T}$, $\mathbf{C}^{0} = \mathbf{T} \cap \mathbf{C}$. *Let* $T_1 < T_2 <$ $< T_k$ *be a topological order of* $G[\mathbf{T}]$ *where* $An(S) < \mathbf{T} \cap An(S)$. *Then, for any cftree T :*

-operator (marginalization): If there is no directed arrow with tail in $\mathbf{C}^{0}$ *and head in* $\mathbf{C} \cup \{S\}$ *in* $G[\mathbf{T} \cup \{S\}]$, $Q[\mathbf{T} \mid S]$ $!$ $Q[\mathbf{C} \mid S]$ *is a valid edge for T with mapping*

$$Q[\mathbf{C} \mid S] = \sum_{\mathbf{c}^{0}} Q[\mathbf{T} \mid S]. \tag{6.35}$$

-operator (recovery): If $\mathbf{C} \setminus An(S) = \emptyset$ *; and there is no bidirected edge between* $\mathbf{C}$ *and* $\mathbf{C}^{0} \cup \{S\}$ *in* $G[\mathbf{T} \cup \{S\}]$, $Q[\mathbf{T} \mid S]$ $!$ $Q[\mathbf{C}]$ *is a valid edge for T and the corresponding mapping is*

$$Q[\mathbf{C}] = \prod_{T_i \in \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \mid S]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \mid S]}. \tag{6.36}$$

^{0}*-operator (recovery residue): If* $\mathbf{C}^{0} \setminus An(S) = \emptyset$ *; and there is no bidirected edge between* $\mathbf{C} \cup \{S\}$ *and* $\mathbf{C}^{0}$ *in* $G[\mathbf{T} \cup \{S\}]$, $Q[\mathbf{T} \mid S]$ $^{0}!$ $Q[\mathbf{C} \mid S]$ *is a valid edge for T the corresponding mapping is*

$$Q[\mathbf{C} \mid S] = Q[\mathbf{T} \mid S] \prod_{T_i \in \mathbf{C}^{0}} \frac{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \mid S]}{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \mid S]}. \tag{6.37}$$

The -operator in lemma 8 is a natural generalization of the -operator seen so far. Specifically, if S is ignored, the operator reduces to the one in theorem 7. The - and ^{0}-operators are specifically tailored for selection biased c-factors (those conditioned on S). Moreover, they are similar to the -operator as they produce c-factors with complementary scopes.

The -operator permits the derivation of an unconditional c-factor from a selection-biased one, as long as the variables in the derived c-factor are not ancestors of S and they are not in the same c-component as those ancestors. The ^{0}-operator is complementary to the -operator as it yields a conditional c-factor that retains the variables that could not

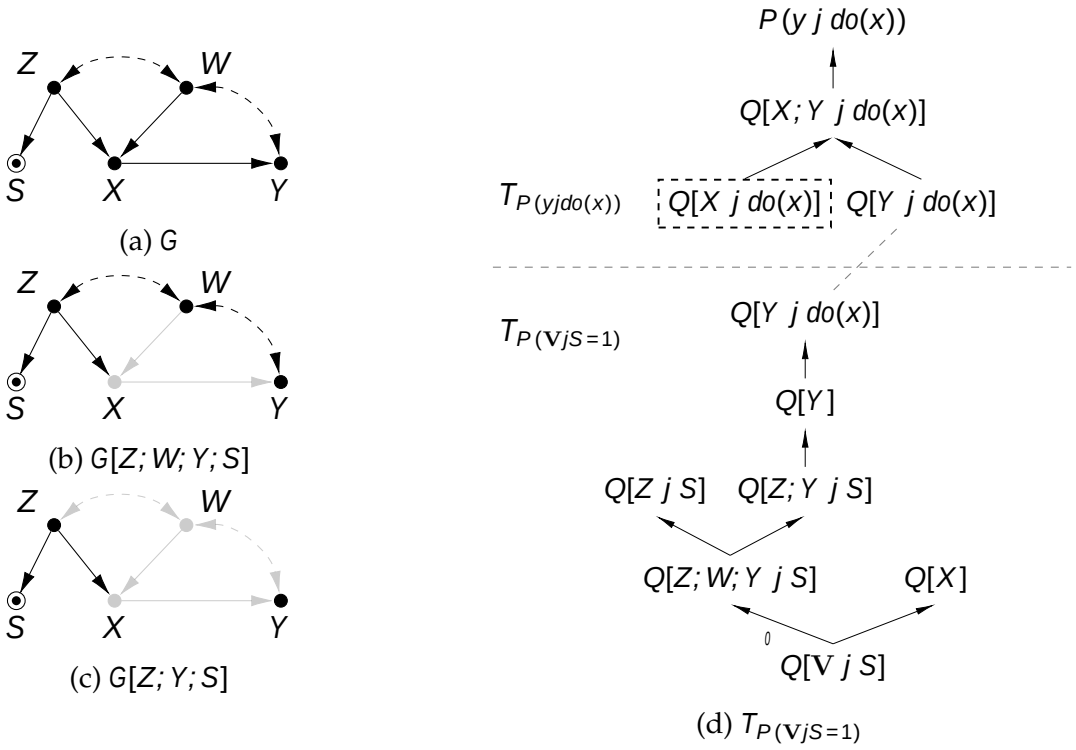


Figure 6.7: A causal diagram and a d-tree for $P(\mathbf{V} \mid S = 1)$

be recovered with the -operator. We give an example of the use of these operators in the construction of a d-tree to recover a causal effect.

Example 37 (Recovering a causal effect from selection bias)**.** Consider the causal diagram in fig. 6.7(a) (previously shown in fig. 6.2(c)) and suppose we are generating a d-tree for $P(\mathbf{V} \mid S = 1)$ where the root is the equivalent $Q[\mathbf{V} \mid S]$. First, we can use the -operator to obtain $Q[\mathbf{X}]$ as X is not an ancestor of S and there is no bidirected arrow between X and any variable in $\{Z; W; Y; S\}$. Also from the root, ^{0}-operator gives $Q[Z; W; Y \mid S]$. From the latter, the -operator allows us to remove W as there is no arrow from W to any other node in $G[Z; W; Y; S]$ (fig. 6.7(b)). Similarly, one can also marginalize Y to obtain $Q[Z \mid S]$. Finally, we can apply the -operator again to derive $Q[Y]$ as $Y \notin An(S)$ in $G[Z; Y; S]$ (fig. 6.7(c)).

Now suppose the query is $P(y \mid do(x))$, then GENQUERYTREE produces the q-tree $T_{P(yjdo(x))}$ in fig. 6.7(d). Then, the c-factor $Q[Y \mid do(x)]$ is mapped from $T_{P(\mathbf{V}jS=1)}$. Following

the path from $T_{P(\mathbf{v}jS=1)}$ to $P(y\ j\ do(x))$, we have

$$Q[Z; W; Y\ j\ S] \overset{0}{=} Q[Z; W; X; Y\ j\ S] \frac{\sum_{x'';y} Q[Z; W; X; Y\ j\ S]}{\sum_{y} Q[Z; W; X; Y\ j\ S]} \tag{6.38}$$

$$= P(\mathbf{v}\ j\ S = 1) \frac{P(z; w\ j\ S = 1)}{P(z; w; x\ j\ S = 1)} \tag{6.39}$$

$$= P(y\ j\ z; w; x; S = 1) P(z; w\ j\ S = 1) \tag{6.40}$$

$$Q[Z; Y\ j\ S] = \sum_{w} Q[Z; W; Y\ j\ S] \tag{6.41}$$

$$= \sum_{w} P(y\ j\ z; w; x; S = 1) P(z; w\ j\ S = 1) \tag{6.42}$$

$$Q[Y] = \frac{Q[Z; Y\ j\ S]}{\sum_{y} Q[Z; Y\ j\ S]} \tag{6.43}$$

$$= \frac{\sum_{w} P(y\ j\ z; w; x; S = 1) P(w; z\ j\ S = 1)}{\sum_{w;y} P(y\ j\ z; w; x; S = 1) P(w; z\ j\ S = 1)} \tag{6.44}$$

$$= \sum_{w} P(y\ j\ z; w; x; S = 1) P(w\ j\ z; S = 1) \tag{6.45}$$

Finally, the query is equal to

$$P(y\ j\ do(x)) = \sum_{x'} Q[X; Y\ j\ do(x)] \tag{6.46}$$

$$\doteq \sum_{x'} 1[x = x'] Q[Y] \tag{6.47}$$

$$= \sum_{w} P(y\ j\ z; w; x; S = 1) P(w\ j\ z; S = 1); \tag{6.48}$$

which is the expression previously given in eq. (6.4).

This extension to d-trees can be easily incorporated in the framework developed so far in a natural way. When running GENINPUTTREE, if the current c-factor is conditioned on S, we consider the $\doteq$-operator in lemma 8 instead of the previous ones. Similarly, instead

of considering the -operator, we apply the or ^{0}-operator depending on whether a set $\mathbf{C}$ containing the variables in the target c-factor satisfies the condition of one or the other operator.

Theorem 18 (-, -, -, ^{0}-, and -operators soundness and completeness for sb-ID)**.** *Given a causal inference task with signature* $I_{sb\text{-}ID} = \langle Q = P(\mathbf{y} \mid do(\mathbf{x})); \{P(\mathbf{V} \mid S = 1)\}; \{G\}\rangle$, *the query* Q *is recoverable from* $P(\mathbf{V} \mid S = 1)$ *and* G *if and only if* C-INFER *finds a mapping using the , , and* ^{0}*-operators. Moreover, the task is decided in* $O(n^2(n + m))$ *time, where* $n = |\mathbf{V}|$ *and* m *is the number of edges in* G.

While theorem 18 characterizes the solution to *sb-ID* tasks instances via C-INFER, there is a simple graphical condition that is necessary for *sb-ID*, shown next.

Theorem 19 (Necessary Graphical Condition for *sb-ID*)**.** *A sb-ID task with signature* $I = \langle P(\mathbf{y} \mid do(\mathbf{x}); S = 1); \{P(\mathbf{V} \mid S = 1)\}; G\rangle$ *is solvable only if* $(\mathbf{Y} \perp\!\!\!\perp S)_{G^{pbd}_{\mathbf{XY}}}$.

The graphical condition in theorem 19 provides a simple way to detect some instances when the causal effect $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable from the selection biased distribution. For instance, consider the causal diagram in fig. 6.8(a) and the corresponding $G^{pbd}_{\mathbf{XY}}$ in fig. 6.8(b). The path $Y \leftarrow Z \rightarrow W \rightarrow S$ witnesses the non-recoverability of $P(y \mid do(x))$ as per theorem 19.

6.4.1 Combining biased data and partial unbiased data

In this section, we consider again the *sbt-ID* task that includes biased data and unbiased data in the form of a distribution $P(\mathbf{T})$, $\mathbf{T} \subseteq \mathbf{V}$, as input. When such external data is available, c-factors of the query that are not recoverable from biased data alone could be computable with the help of the extra dataset.

An additional input distribution could be helpful by providing required c-factors that are not recoverable from the selection biased distribution. This situation is similar to the

sTR task (section 5.4), where we looked at different ways of decomposing the query. The following example illustrates this point.

Example 38 (Mixing c-factors from distributions with and without selection bias)**.** Consider the task with signature $I = \langle P(y \mid do(x)); \mathsf{P}; G \rangle$, where $\mathsf{P} = \{P(\mathbf{V} \mid S = 1); P(Z)\}$ and G is the causal diagram in fig. 6.8(a). Using $G_{\overline{X}}$ (fig. 6.8(c)), the GENQUERYTREE algorithm (algorithm 8) will generate a q-tree T_Q with two branches corresponding to observing $\mathbf{V}$ and $\{Z\}$ among the input distributions.

Next, for each of the input distributions GENINPUTTREE produces $T_{P(\mathbf{V} \mid S=1)}$ and $T_{P(Z)}$ (bottom part of fig. 6.8(d)). Then, MAPFACTORS is able to map $Q[X \mid do(x)]$ (trivially), $Q[Y \mid do(x)]$ and $Q[Z \mid do(x), R]$, as shown with dashed lines in fig. 6.8(d).

Finally, COMPOSEQUERY will use the edges marked in blue in fig. 6.8(d) to produce

$$P(y \mid do(x)) = \sum_{z} P(y \mid x, z, S = 1) P(z). \tag{6.49}$$

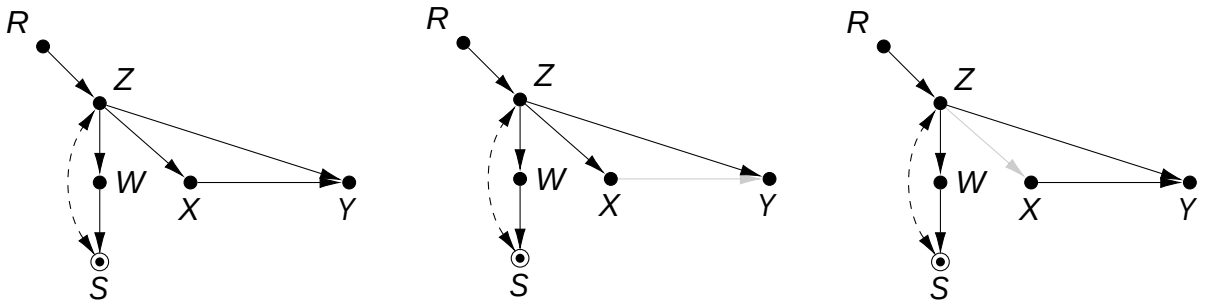
Remarkably, although neither $Q[R \mid do(x)]$ nor $Q[Z \mid do(x)]$ are recoverable individually, $\sum_{R} Q[R \mid do(x)] Q[Z \mid do(x)]$ is equal to the input $P(z)$.

Aside from providing particular c-factors in the q-tree, additional unbiased data can be combined with biased data to compute c-factors that are not computable from any individual source. To formalize this idea, we first introduce a new cftree operator.

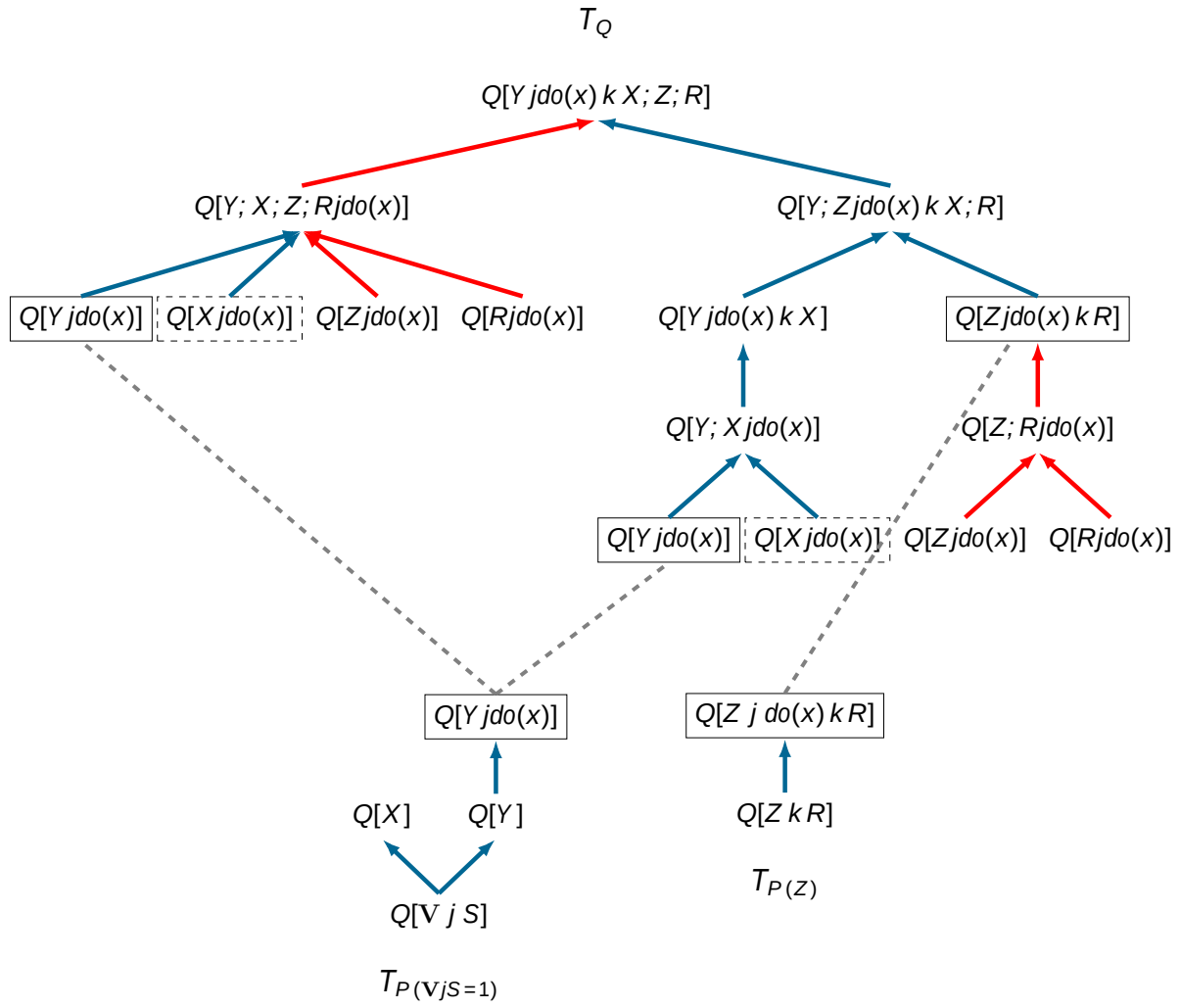
Lemma 9 (- and $^{\emptyset}$-operators for combining c-factors)**.** *Let endogenous sets* $\mathbf{T}, \mathbf{C}, \mathbf{C}^{\emptyset}, \mathbf{Z}, \mathbf{L}$ *be such that* $\mathbf{C} \quad \mathbf{T}$, $\mathbf{C}^{\emptyset} = \mathbf{T} \cap \mathbf{C}$ *and* $\mathbf{L} = An(\mathbf{Z})_{G[\mathbf{T} \cup \{S\}]}$. *Also, let* $\mathbf{A} = An(\mathbf{C}; S) \cap \{S\}$ *in* $G[\mathbf{T} \cup \{S\}]$ *and* $\mathbf{R} = An(S) \cap \{S\}$ *in* $G[\mathbf{T} \cup \{S\}]_{\overline{\mathbf{Z} \setminus \mathbf{T}}}$, $C_1 < C_2 < \cdots$ *be a topological order of* $G[\mathbf{A} \cap \mathbf{R}; \mathbf{Z} \setminus \mathbf{A}]$ *and*

$$Q^{\emptyset} = \frac{\sum_{\mathbf{t} \in \mathbf{a}} Q[\mathbf{T} \mid S] \sum_{\mathbf{z} \in \mathbf{r}} Q[\mathbf{Z} \mid \mathbf{L}]}{\sum_{\mathbf{t} \in \mathbf{r}} Q[\mathbf{T} \mid S]}. \tag{6.50}$$

Then, for any pair $Q[\mathbf{T} \mid S]$ *and* $Q[\mathbf{Z} \mid \mathbf{L}]$*:*



(a) G (b) $G^{pbd}_{\mathbf{XY}}$ (c) $G_{\overline{X}}$



(d) Representation of the matching process of the query and input cftrees

Figure 6.8: Causal diagrams associated with the input distributions and query of an *sbt-ID* task. For each one of the distributions a d-tree is generated.

-operator (unbiasing): If there is no path (regardless of the direction) between $\mathbf{C}$ and S in $G[\mathbf{A} \cup \{S\}]_{\underline{\mathbf{Z} \setminus \mathbf{A}}}$ and there is no bidirected edge between $\mathbf{C}$ and $\mathbf{C}^0 \cup \{S\}$ in $G[\mathbf{T} \cup \{S\}]_{\underline{\mathbf{Z} \setminus \mathbf{T}}}$, then $\{Q[\mathbf{T} \mid S]; Q[\mathbf{Z} \mid \mathbf{L}]\} \to Q[\mathbf{C}]$ are valid edges for a cftree T and the corresponding mapping is

$$Q[\mathbf{C}] = \prod_{C_i \in \mathbf{C}} \frac{\sum_{c_i+1,\ldots,c_k} Q^0}{\sum_{c_i,\ldots,c_k} Q^0}. \tag{6.51}$$

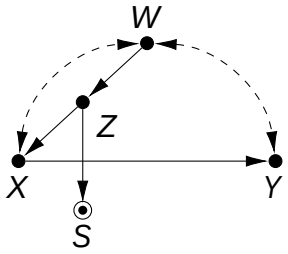
^{0}-operator (unbiasing residue): If $\mathbf{C}^0 \setminus An(S) = \emptyset$, there is no bidirected edge between $\mathbf{C}^0$ and $\mathbf{C} \cup \{S\}$, and there is no bidirected edge between $\mathbf{Z} \setminus \mathbf{R}$ and $(\mathbf{R} \cup \{S\}) \cap \mathbf{Z}$ in $G[\mathbf{T} \cup \{S\}]_{\underline{\mathbf{Z} \setminus \mathbf{T}}}$, then $\{Q[\mathbf{T} \mid S]; Q[\mathbf{Z} \mid \mathbf{L}]\} \to Q[\mathbf{C} \mid S]$ are valid edges for a cftree T and the corresponding mapping is

$$Q[\mathbf{C} \mid S] = Q[\mathbf{T} \mid S] \prod_{C_i \in \mathbf{C}^0} \frac{\sum_{c_i,\ldots,c_k} Q^0}{\sum_{c_i+1,\ldots,c_k} Q^0}. \tag{6.52}$$

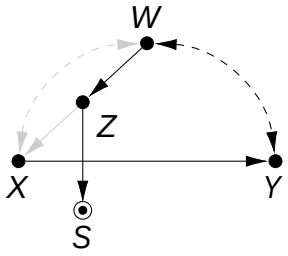
The -operator allows for a c-factor $Q[\mathbf{C}]$ to be computed from a combination of two c-factors coming from different inputs: a selection biased c-factor and a marginal c-factor that is not selection biased. Complementarily, the ^{0}-operator licenses the expansion to a biased c-factor that retaining a set $\mathbf{C}$ not satisfying the -operator. For concreteness, consider the following example.

Example 39 (Combining selection biased distributions and unbiased distributions over subsets of $\mathbf{V}$)**.** Consider a *sbt-ID* task with the casual diagram G shown in fig. 6.9(a), $\mathsf{P} = \{P(\mathbf{V} \mid S = 1); P(Z)\}$ and $Q = P(y \mid do(x))$. Because Y is its own only ancestor in $G_{\overline{X}}$ (fig. 6.9(b)), GENQUERYTREE generates a q-tree $T_{P(y|do(x))}$ (above the dotted line in fig. 6.8(d)) where the root and only c-factor is $Q[Y \mid do(x)]$.

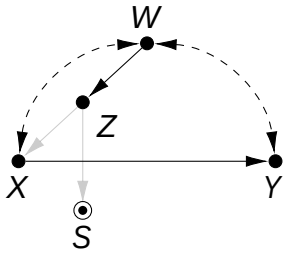
For the d-trees, if we consider $T_{P(\mathbf{V}|S=1)}$ and $T_{P(Z)}$ individually, there are no operators that could lead to $Q[Y \mid do(x)]$. We can, however, resort to the -operator and combine the two distributions as shown in the bottom portion of fig. 6.9(d). Specifically, the rule applies by taking $\mathbf{T} = \mathbf{V}$, $\mathbf{C} = \mathbf{V}$, $\mathbf{C}^0 = \emptyset$, $\mathbf{Z} = \{Z\}$, $\mathbf{L} = \{W\}$, $\mathbf{A} = \mathbf{V}$ and $\mathbf{R} = \{Z\}$. The condition is satisfied because no variable in $\mathbf{V}$ is an ancestor of S and there are no bidirected arrows



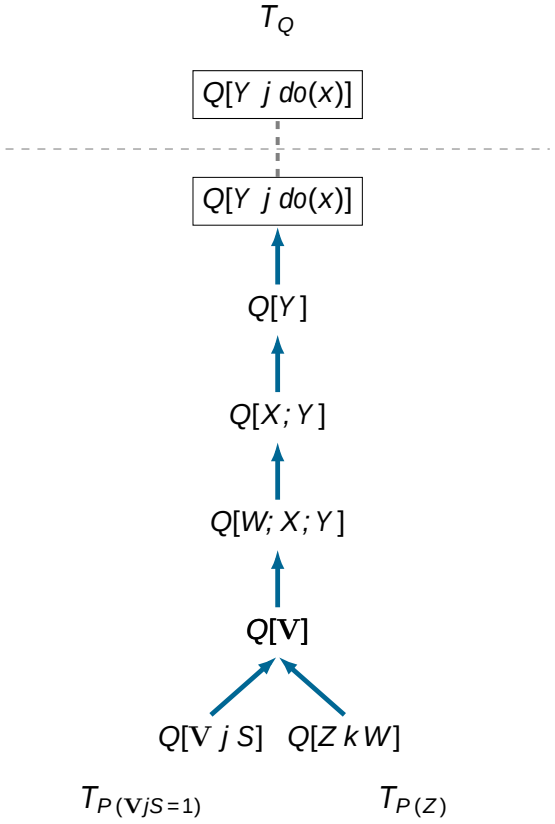
(a) G



(b) $G_{\overline{X}}$



(c) $G_{\underline{Z}}$



(d) Matching between the query and input cftrees

Figure 6.9: Causal diagrams and cftrees for an instance where a selection-biased distribution is combined with an unbiased distribution to recover c-factors that are not computable from any of the individual distributions.

between $\mathbf{V}$ and S in $G_{\underline{Z}}$ (shown in fig. 6.9(c)). Then, it follows that $Q[\mathbf{V}]$ is computable from $P(\mathbf{V}\ j\ S = 1)$ and $P(Z)$ using the -operator.

$$Q[\mathbf{V}] = \frac{Q[\mathbf{V}\ j\ S]Q[Z\ k\ W]}{\sum_{w;x;y} Q[\mathbf{V}\ j\ S]} \tag{6.53}$$

$$\stackrel{\text{def.}}{=} \frac{P(\mathbf{v}\ j\ S = 1)P(z)}{\sum_{w;x;y} P(\mathbf{v}\ j\ S = 1)} \tag{6.54}$$

$$= P(w;x;y\ j\ z;S = 1)P(z): \tag{6.55}$$

From this point, we can derive $Q[Y]$ from $Q[\mathbf{V}] = P(\mathbf{v})$ as with a classical *obs-ID* task.

Example 40 (A more elaborate recoverability task)**.** Consider a *sbt-ID* task with the casual diagram G shown in fig. 6.10(a), $\mathsf{P} = \{P(\mathbf{V} \mid S=1); P(V_2, V_3, V_6)\}$ and $Q = P(y \mid do(x))$.

Because the set of ancestors of Y in $G_{\overline{X}}$ (fig. 6.10(b)) are $\{Y, V_6, V_5\}$, GENQUERYTREE generates a q-tree $T_{P(y|do(x))}$ (above the dotted line in fig. 6.8(d)) with $Q[Y \mid do(x), V_6, V_5]$ as the root.

Individually, none of the input d-trees, $T_{P(\mathbf{V}|S=1)}$ and $T_{P(V_2,V_3,V_5)}$, allow for the derivation of $Q[Y \mid do(x)]$. Similar to the previous example, we will combine them, this time using the $^{\emptyset}$-operator, as shown in the bottom of fig. 6.10(d). Specifically, the rule applies by taking $\mathbf{T} = \mathbf{V}$, $\mathbf{C}^{\emptyset} = \{V_2\}$, $\mathbf{C} = \mathbf{V} \setminus \{V_2\}$, $\mathbf{Z} = \{V_2, V_3, V_6\}$, $\mathbf{L} = \{V_1, V_4, V_5\}$, $\mathbf{A} = \{V_1, V_2, V_7\}$ and $\mathbf{R} = \{V_2, V_7\}$. The condition holds because no variable in $\mathbf{C}^{\emptyset}$ is an ancestor of S and there are no bidirected arrows between $\mathbf{C}^{\emptyset} \cup \{S\}$ and $\mathbf{C}$ in $G_{\underline{V_2,V_3,V_6}}$ (shown in fig. 6.10(c)). Then, it follows that $Q[\mathbf{V} \setminus \{V_2\} \mid S]$ is computable from $P(\mathbf{V} \mid S=1)$ and $P(V_2, V_3, V_6)$. The corresponding mapping is

$$Q^{\emptyset} = \frac{\sum_{\mathbf{v} \setminus \{v_1, v_2, v_7\}} Q[\mathbf{V} \mid S] \sum_{v_3, v_6} Q[V_2, V_3, V_6 \mid V_1, V_4, V_5]}{\sum_{\mathbf{v} \setminus \{v_2, v_7\}} Q[\mathbf{V} \mid S]} \tag{6.56}$$

$$= \frac{P(v_1, v_2, v_7 \mid S=1) P(v_2)}{P(v_2, v_7 \mid S=1)} \tag{6.57}$$

$$= P(v_1 \mid v_2, v_7, S=1) P(v_2) \tag{6.58}$$

$$Q[\mathbf{V} \setminus \{V_2\} \mid S] \overset{\emptyset}{=} Q[\mathbf{V} \mid S] \frac{\sum_{v_2} Q^{\emptyset}}{Q^{\emptyset}} \tag{6.59}$$

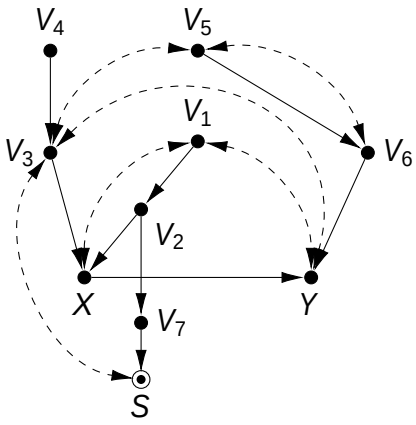
$$\overset{\text{def.}}{=} P(\mathbf{v} \mid S=1) \frac{\sum_{v_2} P(v_1 \mid v_2, v_7, S=1) P(v_2)}{P(v_1 \mid v_2, v_7, S=1) P(v_2)}. \tag{6.60}$$

Next, from $Q[\mathbf{V} \setminus \{V_2\} \mid S]$ we can use the $\sum$-operator to marginalize V_1 and obtain $Q[\mathbf{V} \setminus \{V_1, V_2\} \mid S]$. The $^{\emptyset}$-operator allows the removal of X as X is not an ancestor of S and there are no bidirected arrows between X and $\mathbf{V} \setminus \{V_1, V_2\}$, then we obtain $Q[\mathbf{V} \setminus \{V_1, V_2, X\} \mid S]$. Once again, the $\sum$-operator is used to obtain $Q[Y, V_6, V_5, V_7 \mid S]$. Following this, the $\sum$-operator allows for the recovery of $Q[Y, V_6, V_5]$ as the variables in this

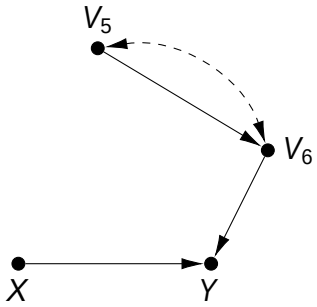
factor are not ancestors of S and there are no bidirected arrows between $\{Y, V_6, V_5\}$ and $\{V_7, S\}$. Finally, the and -operator allows us to derive the target factors $Q[V_6, V_5 \mid do(x)]$ and $Q[Y \mid do(x)]$.

After mapping the factors, COMPOSEQUERY returns the query in terms of $Q[\mathbf{V} \cap \{V_2\} \mid S]$.

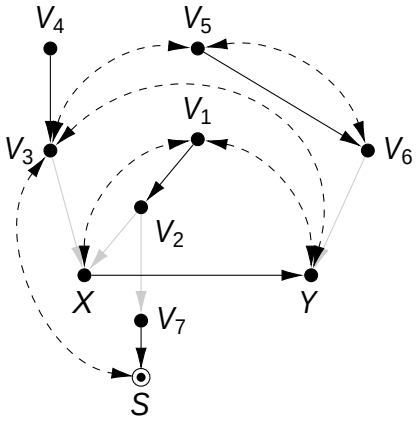
Theorem 20 (-, -, -, ^{0}-, -, ^{0}-, and -operators soundness *sbt-ID*)**.** *Given a causal inference task with signature* $I_{sbt\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})), \{P(\mathbf{V} \mid S = 1), P(\mathbf{T})\}, \{G\}\rangle$*, the query is recoverable from* $P(\mathbf{V} \mid S = 1)$, $P(\mathbf{T})$ *and* G *if* C-INFER *finds a mapping using the* , , , ^{0} *and* -*operators. Moreover, the process takes* $O(n^2(n + m))$ *time, where* $n = |\mathbf{V}|$ *and* m *is the number of edges in* G*.*



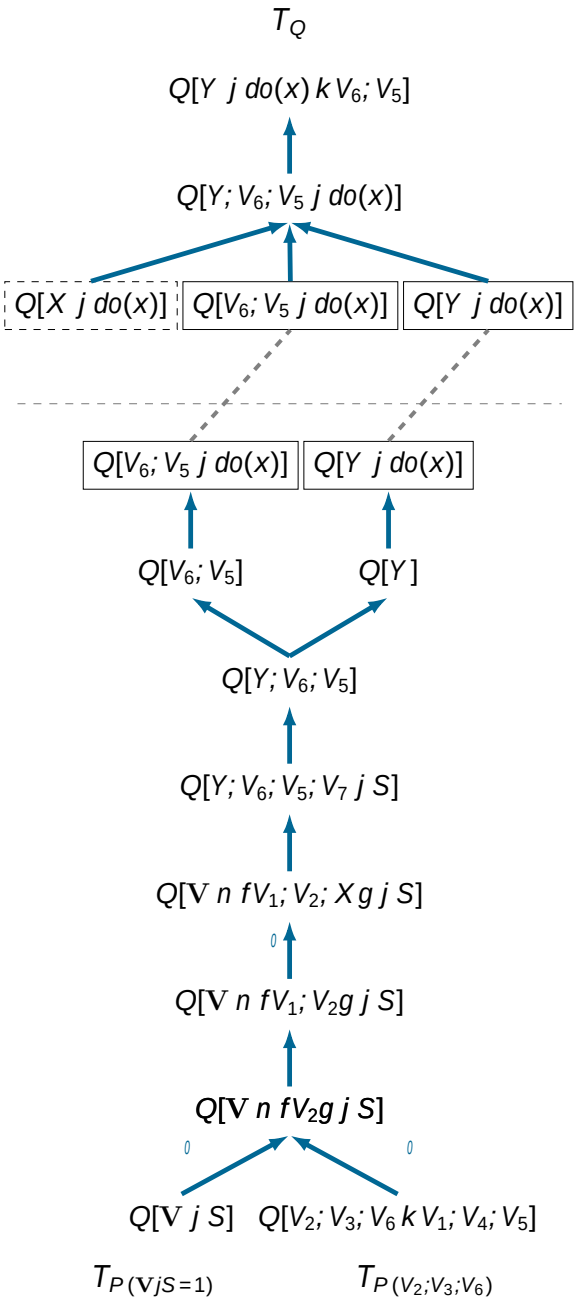
(a) G



(b) $G_{\overline{X}}[An(Y)]$



(c) $G_{\underline{V_2;V_3;V_6}}$



(d) Matching between the query and input cftrees

Figure 6.10: Causal diagrams and cftrees for an instance where a selection-biased distribution is combined with an unbiased distribution to recover c-factors that are not computable from any of the individual distributions.

Chapter 7: Towards Causal Data Fusion

This chapter summarizes the results discussed in this dissertation, including the tasks (section 7.1) and the cftree operators defined within the algorithmic framework (section 7.2). We put these results together in section 7.3 and introduce the most general version of the C-INFER algorithm. In section 7.4, we prove different properties of C-INFER algorithm and exemplify how it can be used to solve tasks involving multiple data fusion dimensions, which goes beyond the results previously discussed. Finally, in section 7.5, we provide a short discussion on different data fusion challenges that are not yet fully accounted for in our framework and deserve further investigation.

7.1 Summary of the Tasks

In chapter 3, we discussed three tasks previously solved in the literature in terms of their cftrees and within the C-INFER framework. Moreover, we proved the soundness, completeness, and efficiency of the framework to solve these tasks (theorems 1 to 3). This was summarized in table 3.1.

Later on, we formalized and proved new soundness and completeness results for tasks in chapters 4 to 6 (theorems 6, 8, 18 and 20). These tasks are summarized in table 7.1 where we list each tasks' name, signature — corresponding query Q, input distributions P, and assumptions $\mathcal{G}$ — the corresponding operators needed to evaluate the corresponding cftrees, and whether the final results are sufficient and necessary.

	Name	Q	P	G	Ops	S&N
-TR	Transportability of soft interventions	$P(\mathbf{y};\ \mathbf{x})$	$fP^i(\mathbf{V};\ \mathbf{z}_j)j\mathbf{Z}_j 2 Z^i g_{Z^i 2 Z}$	G^{Δ}	, , ,	,
s-TR	Statistical transportability	$P(\mathbf{y}j\mathbf{x})$	$P(\mathbf{V}); P(\mathbf{W})$	G^{Δ}	, , ,	,
sb-ID	Recovering from Selection Bias	$P(\mathbf{y}jdo(\mathbf{x}))$	$P(\mathbf{V}jS=1)$	G	, , , 0 ,	,
sbt-ID	Recovering from Selection Bias and external data	$P(\mathbf{y}jdo(\mathbf{x}))$	$P(\mathbf{V}jS=1); P(\mathbf{T})$	G	, , , 0 , , 0 ,	)
exp-TR	Generalizing experimental findings	$P(\mathbf{y}jdo(\mathbf{x}))$	$P(\mathbf{V}jdo(\mathbf{x}); S=1); P(\mathbf{W})$	G^{Δ}	, , , 0 , , , 0 ,	)

Table 7.1: Summary of the causal inference tasks discussed in the dissertation. Each row describes a task with their name, query Q, input distributions P, assumptions in the form of causal diagrams G , the cftree operators used to solve it (Ops), and whether the framework was proven sufficient (,) or sufficient and necessary ( )) for the task.

7.2 Summary of the Cftree Operators

The framework developed throughout the dissertation relies on a set of operators for expanding nodes in cftrees. Table 7.2 lists all the operators defined with their symbol, name, conditions, and edges they license in a cftree. The conditions in the table assume a c-factor with endogenous variables $\mathbf{T}$ $\mathbf{V}$ is to be expanded and that the c-factor being derived contains a set of variables $\mathbf{C}$ $\mathbf{T}$. Depending on the variations, the sets $\mathbf{L}$ and $\mathbf{H}$ are used to represent variables being marginalized in the c-factors and S represents the sampling-selection mechanism for c-factors conditioned on S.

As each operator licenses an edge connecting a pair of c-factors in a cftree, they are associated with a mapping involving one or more factors. The mapping corresponding to each operator is listed in table 7.3.

Sym	Name	Condition	Cftree edge(s)
	Marginalization	No directed arrow from $\mathbf{C}'$ to $\mathbf{C}$ in $G[\mathbf{T}]$	$Q[\mathbf{T}] \;!\; Q[\mathbf{C}]$
	Independence	No bidirected arrow between $\mathbf{C}'$ in $\mathbf{C}$ in $G[\mathbf{T}]$	$Q[\mathbf{T}] \;!\; Q[\mathbf{C}]$ $\{Q[\mathbf{T}] \quad Q[\mathbf{C}_i]\}_{i=1,\ldots,k}$
	Regime Invariance	$\mathbf{T} \setminus \mathbf{X} = \emptyset$	$Q[\mathbf{T} \mid \pi_x] \quad Q[\mathbf{T}]$ $Q[\mathbf{T} \mid \pi_x] \;!\; Q[\mathbf{T}]$
	Domain Invariance	$\mathbf{T} \setminus \pi^{a,b} = \emptyset$	$Q^a[\mathbf{T}] \quad Q^b[\mathbf{T}]$ $Q^a[\mathbf{T}] \;!\; Q^b[\mathbf{T}]$
	Unconditioning	$\mathbf{T}_1 \quad \mathbf{C} \quad \mathbf{T}$ and there is no path between $\mathbf{C} \cup \mathbf{H}$ and $\mathbf{C}' \cup \mathbf{H}'$ in $G[\mathbf{T}, \mathbf{L}]_{\underline{\mathbf{T}_2}}$	$Q[\mathbf{T}_1 \mid \mathbf{T}_2, \mathbf{L}] \quad Q[\mathbf{C}, \mathbf{H}]$
	Recovery	$\mathbf{C} \setminus An(S) = \emptyset$ and no bidirected arrow between $\mathbf{C}$ and $\mathbf{C}' \cup S$ in $G[\mathbf{T} \cup S]$	$Q[\mathbf{T} \mid S] \;!\; Q[\mathbf{C}]$
$'$	Recovery residue	$\mathbf{C}' \setminus An(S) = \emptyset$ and no bidirected arrow between $\mathbf{C} \cup S$ and $\mathbf{C}'$ in $G[\mathbf{T} \cup S]$	$Q[\mathbf{T} \mid S] \;\overset{'}{!}\; Q[\mathbf{C} \mid S]$
	Unbiasing	$\mathbf{C} = An(\mathbf{C})$, $\mathbf{C} \setminus An(S) = \emptyset$ and no bidirected arrow between $\mathbf{C}$ and $\mathbf{C}' \cup S$, all in $G[\mathbf{T} \cup S]_{\underline{\mathbf{Z} \setminus \mathbf{T}}}$	$\{Q[\mathbf{T} \mid S], Q[\mathbf{Z}, \mathbf{L}]\} \;!\; Q[\mathbf{C}]$
$'$	Unbiasing residue	$\mathbf{C}' = An(\mathbf{C}')$, $\mathbf{C}' \setminus An(S) = \emptyset$ and no bidirected arrow between $\mathbf{C}'$ and $\mathbf{C} \cup S$, all in $G[\mathbf{T} \cup S]_{\underline{\mathbf{Z} \setminus \mathbf{T}}}$	$\{Q[\mathbf{T} \mid S], Q[\mathbf{Z}, \mathbf{L}]\} \;\overset{'}{!}\; Q[\mathbf{C} \mid S]$

Table 7.2: Summary of the cftree operators defined throughout the thesis.

Sym	Cftree edge(s)	Mapping
	$Q[\mathbf{T}] \;!\; Q[\mathbf{C}]$	$Q[\mathbf{C}] = \sum_{\mathbf{c}'} Q[\mathbf{T}]$
	$Q[\mathbf{T}] \;!\; Q[\mathbf{C}]$	$Q[\mathbf{C}] = \prod_{T_i \supset \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T}]}$
	$\{Q[\mathbf{T}] \quad Q[\mathbf{C}_i]\}_{i=1,\ldots,k}$	$Q[\mathbf{T}] = \prod_{i=1}^{k} Q[\mathbf{C}_i]$
	$Q[\mathbf{T} \mid \mathbf{x}] \quad Q[\mathbf{T}]$ $Q[\mathbf{T} \mid \mathbf{x}] \;!\; Q[\mathbf{T}]$	$Q[\mathbf{T}] = Q[\mathbf{T} \mid \mathbf{x}]$
	$Q^a[\mathbf{T}] \quad Q^b[\mathbf{T}]$ $Q^a[\mathbf{T}] \;!\; Q^b[\mathbf{T}]$	$Q^a[\mathbf{T}] = Q^b[\mathbf{T}]$
	$Q[\mathbf{T}_1 \mid \mathbf{T}_2 \wedge \mathbf{L}] \quad Q[\mathbf{C} \wedge \mathbf{H}]$	$Q[\mathbf{T}_1 \mid \mathbf{T}_2 \wedge \mathbf{L}] = \frac{\sum Q[\mathbf{C} \wedge \mathbf{H}]}{\sum_{\mathbf{t}_1} Q[\mathbf{C} \wedge \mathbf{H}]}$
	$Q[\mathbf{T} \mid S] \;!\; Q[\mathbf{C}]$	$Q[\mathbf{C}] = \prod_{T_i \supset \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \mid S]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \mid S]}$
$'$	$Q[\mathbf{T} \mid S] \;'!\; Q[\mathbf{C} \mid S]$	$Q[\mathbf{C} \mid S] = Q[\mathbf{T} \mid S] \prod_{T_i \supset \mathbf{C}'} \frac{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \mid S]}{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \mid S]}$
	$\{Q[\mathbf{T} \mid S], Q[\mathbf{Z} \wedge \mathbf{L}]\} \;'!\; Q[\mathbf{C} \mid S]$	$Q[\mathbf{C}] = \frac{\sum_{\mathbf{t} \in (\mathbf{c} \cap \mathbf{r})} Q[\mathbf{T} \mid S] \sum_{\mathbf{z} \in \mathbf{c}} Q[\mathbf{Z} \wedge \mathbf{L}]}{\sum_{\mathbf{t} \in ((\mathbf{c} \setminus \mathbf{z}) \cap \mathbf{r})} Q[\mathbf{T} \mid S]}$
$'$	$\{Q[\mathbf{T} \mid S], Q[\mathbf{Z} \wedge \mathbf{L}]\} \;'!\; Q[\mathbf{C} \mid S]$	$Q[\mathbf{C} \mid S] = Q[\mathbf{T} \mid S] \frac{\sum_{\mathbf{t} \in ((\mathbf{c}' \setminus \mathbf{z}) \cap \mathbf{r})} Q[\mathbf{T} \mid S]}{\sum_{\mathbf{t} \in (\mathbf{c}' \cap \mathbf{r})} Q[\mathbf{T} \mid S]} \frac{\sum_{\mathbf{z} \in \mathbf{c}'} Q[\mathbf{Z} \wedge \mathbf{L}]}{\sum_{\mathbf{z} \in \mathbf{c}'} Q[\mathbf{Z} \wedge \mathbf{L}]}$

Table 7.3: Summary of the mapping corresponding to each cftree edge as defined by the cftree operators.

7.3 Algorithm

The C-INFER algorithm has been extended and refined throughout this dissertation to account for the idiosyncrasies of each of the tasks investigated. Combining the developments and operators discussed in chapters 3 to 6, we state its most general versions in algorithms 14 to 18 . Notice that in GENINPUTTREE, $Q[\mathbf{C}]$ is used to represent a target c-factor that could have a specific intervention, domain, or marginalized variables. For simplicity, we write $Q[\mathbf{C}]$ instead of $Q^{a}[\mathbf{C} \,k\, \mathbf{H};\ \mathbf{x}]$.

Algorithm 14 C-INFER($I = \langle Q; \mathsf{P}; \mathcal{G} \rangle$)

Input: A causal inference task I consisting of a query Q, a set of input distributions P, and a set of causal diagrams $\mathcal{G}$ over observable variables $\mathbf{V}$.
Output: Q as a function of P or FAIL.

Let T_Q GENQUERYTREE(I).
for each $P \in \mathsf{P}$ **do**
  Let T_P GENINPUTTREE($P; I; T_Q$).
end for
  MAPFACTORS($T_Q; \{T_P\}_{P \in \mathsf{P}}$).
if there are not enough $Q[\mathbf{D}_j] \in T_Q$ such that $(Q[\mathbf{D}_j]) = ;$ **then**
  return FAIL.
else
  return COMPOSEQUERY($T_Q;\ $).
end if

7.4 Data Fusion Tasks

Although we characterized properties of the tasks listed in tables 3.1 and 7.1, C-INFER and the operators in table 7.2 can be used for a variety of other tasks. For instance, most of the tasks in the table have marginal distributions as queries except for *s-TR* for which we used the operator to rewrite the conditional query as a marginal one (i.e., the c-favor analogous to the definition of conditional probability). Further building on the operator, we note that our framework is both sufficient and necessary for solving conditional queries for the tasks of *obs-ID*, *g-ID*, *g-TR*, and *-TR*.

Algorithm 15 GENINPUTTREE(P; $I = \langle Q; \mathbf{P}; G\ i; T_Q$)

Input: A distribution P (could be interventional, from a particular domain, have selection bias and be over a subset of $\mathbf{V}$), a causal inference task I and a q-tree T_Q.
Output: A d-tree T for P.

1: Let $G_P \in \mathcal{G}$ be the causal diagram corresponding to P.
2: Let $\mathbf{W} \subseteq \mathbf{V}$ be the scope of P
3: Initialize T with root $Q[\mathbf{W}] \cup An(\mathbf{W})_{G_P}]$ (or $Q[\mathbf{W} \mid S] \cup An(\mathbf{W})_{G_P}]$ if P is selection biased).
4: **for** each target $Q[\mathbf{C}] \in T_Q$, starting from the root, at every $Q[\mathbf{T}]$ generated in T **do**
5: **if** If $\mathbf{C}$ is intervened or has discrepancies in $Q[\mathbf{T}]$ **then** give up on $Q[\mathbf{C}]$.
6: **if** $\mathbf{T} = \mathbf{C}$ and model and domain match **then** move to next $Q[\mathbf{C}]$
7: **if** $\mathbf{T} = \mathbf{C}$ and the models does not match **then** expand $Q[\mathbf{T}] \to Q[\mathbf{C}]$ with target intervention.
8: **if** $\mathbf{T} = \mathbf{C}$ **then** expand $Q[\mathbf{T}] \to Q[\mathbf{C}]$ with target domain.
9: Let $\mathbf{A} \leftarrow An(\mathbf{C})_{G[\mathbf{T}]}$.
10: **if** $\mathbf{A} \subsetneq \mathbf{T}$ **then** expand $Q[\mathbf{T}] \to Q[\mathbf{A}]$.
11: **if** $Q[\mathbf{T}]$ is not selection biased **then**
12: **if** $G[\mathbf{T}]$ has more than one (marginalized) c-component **then** expand $Q[\mathbf{T}] \to Q[\mathbf{W}]$, where $\mathbf{W}$ is the union of the c-components intersecting $\mathbf{C}$.
13: **else**
14: **if** $\mathbf{C}$ is a subset of the union of c-components of $G[\mathbf{T}]$ with no $An(S)_{G[\mathbf{T}]}$ **then** expand $Q[\mathbf{T}] \to Q[\mathbf{W}]$, where $\mathbf{W}$ is the union of the c-components intersecting $\mathbf{C}$.
15: **if** $\mathbf{C}$ is a subset of the union of c-components of $G[\mathbf{T}]$ intersecting some $An(S)_{G[\mathbf{T}]}$ **then** expand $Q[\mathbf{T}] \xrightarrow{0} Q[\mathbf{W} \mid S]$, where $\mathbf{W}$ is the union of the c-components intersecting $An(S)_{G[\mathbf{T}]}$.
16: **end if**
17: Give up on $Q[\mathbf{C}]$. . No operator left
18: **end for**

Theorem 21 (C-INFER soundness and completeness w.r.t. solving conditional *obs-ID*, *g-ID*, *g-TR*, and *-TR*.)**.** *Consider a causal inference task with one of the following signatures:*

$$I_{\text{obs-IDC}} = \langle P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}); \{P(\mathbf{V})\}; \{G\}\rangle; \tag{7.1}$$

$$I_{\text{g-IDC}} = \langle P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}); \mathbf{P}; \{G\}\rangle; \tag{7.2}$$

$$I_{\text{g-TRC}} = \langle P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}); \mathbf{P}; G^{\Delta}\rangle; \text{ or} \tag{7.3}$$

$$I_{\text{-TRC}} = \langle P(\mathbf{y} \mid \mathbf{z}; \mathbf{x}); \mathbf{P}; G^{\Delta}\rangle; \tag{7.4}$$

Algorithm 16 GENQUERYTREE($I = \langle Q; \mathsf{P}; G\ i\rangle$)

Input: A causal inference task such that $Q = P(\mathbf{y} \mid \mathbf{z})$ ($\mathbf{z}$ could be empty) and there exists $G \in \mathcal{G}$ describing the SCM associated with Q. (Q could be interventional in which case G represents an intervention graph.)
Output: T_Q, a q-tree for Q.

1: Let $G_Q \in \mathcal{G}$ be the causal diagram associated with the query.
2: Let $\mathbf{D} \leftarrow An(\mathbf{Y})_{G_Q}$.
3: Initialize T_Q with Q at the root.
4: Expand $Q = Q[\mathbf{Y} \mid \mathbf{Z}] \leftarrow Q[\mathbf{Y} \mid \mathbf{Z}, \mathbf{D}]$ where $(\mathbf{A}, \mathbf{H})$ is the union of the marginalized c-components of $G_Q[\mathbf{D}]$, summing $\mathbf{D} \cap \mathbf{Z}$, that intersects with $\mathbf{Y}$.
5: **for each** distinct $\mathbf{W} \subset \mathbf{V}$ such that some $P \in \mathsf{P}$ is over $\mathbf{W}$ **do**
6: Expand $Q[\mathbf{A}, \mathbf{H}] \leftarrow Q[\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W}), \mathbf{H} \cap \mathbf{W}]$.
7: Expand $Q[\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W}), \mathbf{H} \cap \mathbf{W}] \leftarrow Q[\mathbf{A}_i, \mathbf{H}_i]$ for each marginalized c-component of $G[\mathbf{A} \cup \mathbf{H}]_{\underline{\mathbf{A} \cup (\mathbf{H} \setminus \mathbf{W})}}$.
8: Expand each $Q[\mathbf{A}_i, \mathbf{H}_i] \leftarrow Q[\mathbf{A}_i; \mathbf{H}_i]$
9: Expand each $Q[\mathbf{A}_i; \mathbf{H}_i] \leftarrow Q[\mathbf{C}_i]$ for each c-component of $G[\mathbf{A}_i \cup \mathbf{H}_i]$.
10: **end for**
11: **return** T_Q

where P *matches the unconditional version of the tasks. Then, the query* Q *is identifiable / transportable from* P *and* G^{Δ} *if and only if* C-INFER *finds a mapping using the* $\ ,\ ,\ ,\ ,$ *and* $\ $ *operators. Moreover, the task is decided in* $O(n^2(n+m)p)$ *time, where* $n = |\mathbf{V}|$, m *is the number of edges in* G *and* $p = |\mathsf{P}|$.

In general, the $\ $ operator can be used to handle a conditional query in the context of any task, where the query can be written in terms of a conditional c-factor. Hence, the soundness of C-INFER and the cftree operators entail a sound algorithm for the conditional versions of the *sb-ID*, *sbt-ID*, and *exp-TR* tasks.

Moreover, C-INFER can be seen as an inference engine that accepts arbitrary causal inference tasks beyond those specifically named in this thesis. The algorithm will use all the available operators to search for a solution. Consider the following example to illustrate this point.

Example 41 (Arbitrary Causal Fusion Task)**.** Consider a causal inference task with the

Algorithm 17 MAPFACTORS(T_Q; $\{T_P\}_{P \in \mathbf{P}}$)

Input: A q-tree for Q and one d-tree for each input distribution.
Output: $\Psi : Q[\mathbf{D}_i] \to T$, a mapping from a node $Q[\mathbf{D}_i] \in T_Q$ to a d-tree T such that $Q[\mathbf{D}_i] \in T$.

1: **for each** $Q[\mathbf{Z}, \mathbf{x}] \in T_Q, \mathbf{Z} \subseteq \mathbf{X}$ **do**
2: Assign $\Psi(Q[\mathbf{Z}, \mathbf{x}])$ according to $\mathbf{x}$.
3: **end for**
4: **for each** $Q[\mathbf{D}_j, \mathbf{x}] \in T_Q, \mathbf{D}_j \setminus \mathbf{X} = \emptyset$ **do**
5: **if** $Q[\mathbf{D}_j, \mathbf{x}] \in T_P$ for some $P \in \mathbf{P}$ **then**
6: Assign $\Psi(Q[\mathbf{D}_j, \mathbf{x}]) = T_P$.
7: **end if**
8: **end for**
9: **return** Ψ.

Algorithm 18 COMPOSEQUERY(T_Q; Ψ)

Input: A q-tree for Q and a mapping of nodes in T_Q to nodes in some d-tree T.
Output: A expression for Q.

1: **for each** leaf $Q[\mathbf{C}_i] \in T_Q$ **do**
2: Compute $Q[\mathbf{C}_i]$ as a function of the root of $\Psi(Q[\mathbf{C}_i])$ given by the composition of the functions along the path from the root to $Q[\mathbf{C}_i]$ in that d-tree.
3: **end for**
4: **return** the root of T_Q as a function of the leaves.

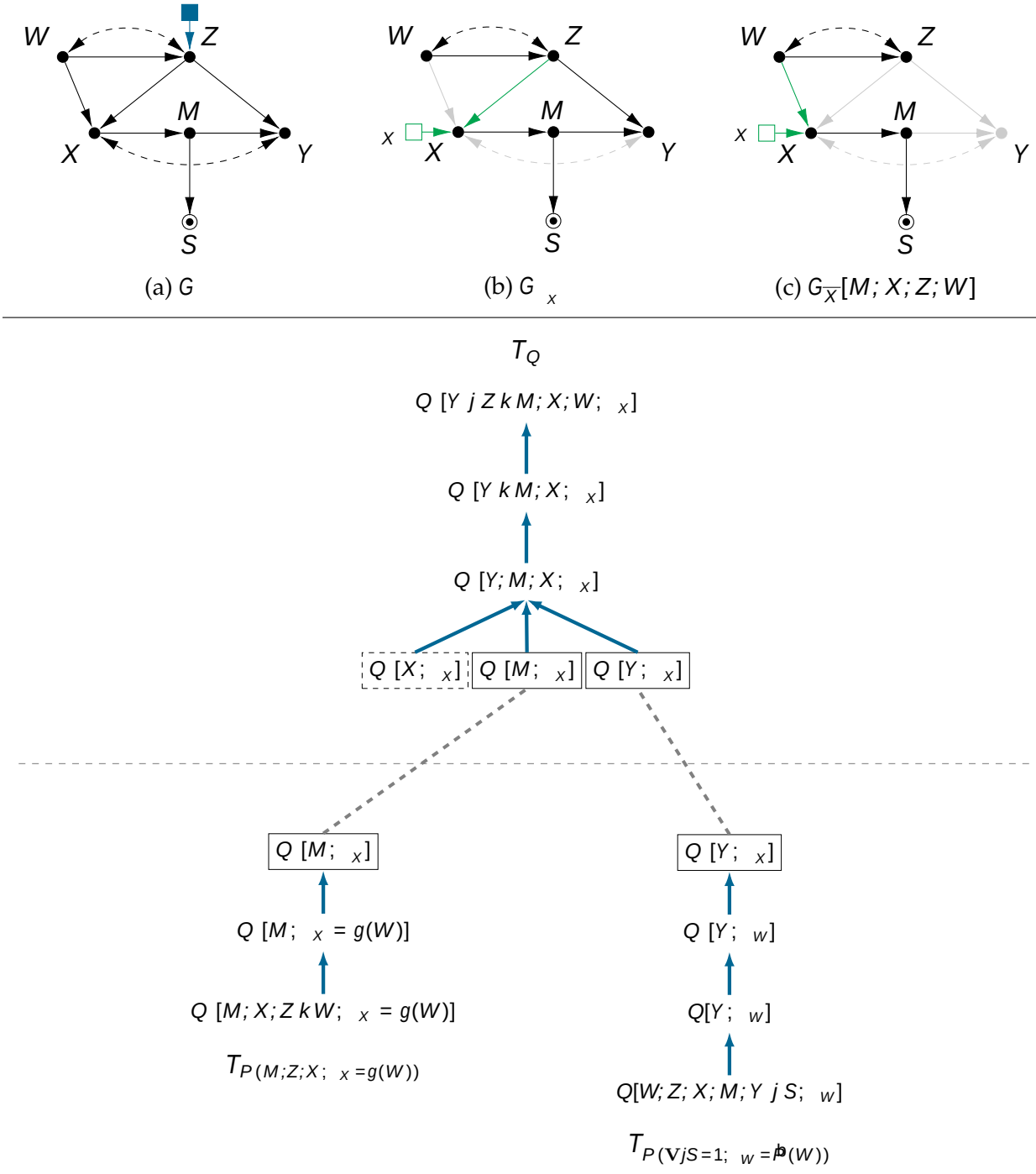
following signature:

$$I = \langle P(y \mid z, \pi_X);\ f_P(M; Z; X, \pi_X); P(W; Z; X; M; Y \mid S = 1, \pi_W)\rangle;\ G^{\Delta}{}_S\rangle; \quad (7.5)$$

where $\pi_X = \langle P(X \mid Z)\rangle$, $\pi_X = \langle g(W)\rangle$, $\pi_W = \langle P(W)\rangle$ and $G^{\Delta}{}_S$ is the selection diagram shown in fig. 7.1(a). Note that this does not match the *sbt-ID*, *s-TR* or *-TR* tasks as the input distributions $\mathbf{P}$ entails several dimensions such as observing only a subset of the endogenous variables, selection bias, and soft experiments. Moreover, the query Q is the conditional causal effect of a soft intervention.

Still, C-INFER can be used to solve a task with this signature. Specifically, we start by running GENQUERYTREE to generate $T_{P(y|z, \pi_X)}$ as shown in the top part of fig. 7.1(d). Then, GENINPUTTREE generates the d-trees $T_{P(M;Z;X, \pi_X)}$ and $T_{P(W;Z;X;Y|S=1, \pi_W)}$ shown in the bottom portion of fig. 7.1(d).

MAPFACTORS matches the c-factors in the query with c-factors in the d-trees, as indi-



(d) Matching between the query and input cftrees

Figure 7.1: Causal diagrams and cftrees for an arbitrary causal inference task covering several data fusion dimensions.

cated by the dashed lines in fig. 7.1(d). Also, the c-factor $Q\ [X;\ \ _X]$ is directly mapped, according to $\ _X$, as $\hat{P}(x\ j\ z)$. COMPOSEQUERY uses the paths (in blue) going from the root of the input d-trees $T_{P(\mathbf{v}jS=1;\ \ W)}$ and $T_{P(M;Z;X;\ \ X)}$ to the root of the q-tree T_Q and the mappings in table 7.3 to write an expression for Q. For simplicity let $\ _X = \hat{P}(X\ j\ Z)$, $\ _X = g(W)$ and $\ _W = \hat{P}(W)$ be written simply as $\ _X$, $\ _X$ and $\ _W$, then we have

$$P\ (y\ j\ z;\ \ _X) = Q\ [Y\ j\ Z\ k\ M;X;W;\ \ _X] \tag{7.6}$$

$$= \frac{Q\ [Y\ k\ M;X;\ \ _X]}{\sum_y Q\ [Y\ k\ M;X;\ \ _X]} \tag{7.7}$$

$$= Q\ [Y\ k\ M;X;\ \ _X]; \tag{7.8}$$

$$Q\ [Y\ k\ M;X;\ \ _X] = \sum_{m;x} Q\ [Y;M;X;\ \ _X]; \tag{7.9}$$

$$Q\ [Y;M;X;\ \ _X] = Q\ [X;\ \ _X]Q\ [M;\ \ _X]Q\ [Y;\ \ _X]; \tag{7.10}$$

$$Q\ [X;\ \ _X] \stackrel{def:}{=} \hat{P}(x\ j\ z); \tag{7.11}$$

$$Q\ [M;\ \ _X] = Q\ [M;\ \ _X] \tag{7.12}$$

$$= \frac{Q\ [M;X;Z;\ \ _X]}{\sum_m Q\ [M;X;Z;\ \ _X]} \tag{7.13}$$

$$= \frac{P(m;z;x;\ \ _X)}{P(z;x;\ \ _X)} \tag{7.14}$$

$$= P(m\ j\ z;x;\ \ _X); \tag{7.15}$$

$$Q\ [Y;\ \ _X] = Q\ [Y;\ \ _W] \tag{7.16}$$

$$= Q[Y;\ \ _W] \tag{7.17}$$

$$= \frac{Q[W;Z;X;M;Y\ j\ S;\ \ _W]}{\sum_y Q[W;Z;X;M;Y\ j\ S;\ \ _W]} \tag{7.18}$$

$$= \frac{P(w;z;x;m;y\ j\ S=1;\ \ _W)}{\sum_y P(w;z;x;m;y\ j\ S=1;\ \ _W)} \tag{7.19}$$

$$= P(y\ j\ w;z;x;m;S=1;\ \ _W): \tag{7.20}$$

Finally, putting it all together leads to the fusion expression:

$$P\ (y\ j\ z;\ \ _X) = \sum_{m;x} \hat{P}(x\ j\ z)P(m\ j\ z;x;\ \ _X)P(y\ j\ w;z;x;m;S=1;\ \ _W); \tag{7.21}$$

which is a function of the input distributions given in the task.

Overall, C-INFER and the cftree operators provide a well-contained —and relatively simple— method to deal with a myriad of problems in causal inference and data fusion. Thanks to a clear definition of the causal inference task, many problems in causal inference and data fusion can be solved by constructing and connecting the cftrees grown with the tree operators used in a clever order. In technical terms, we have shown the sufficiency of this method as well as its completeness for several tasks. The characterizations and algorithms corresponding to the tasks in table 7.1 are completely novel results but for the soundness of *sb-ID*, first shown in [27].

Theorem 22 (C-INFER soundness)**.** *Given a causal inference task with signature $I = \langle Q; \mathbf{P}; G \rangle$, where the query is a conditional associational or interventional query, each input $P \in \mathbf{P}$ is an observational/interventional and partially observed/selection biased distribution, and G is one or more causal diagrams or selection diagrams; then Q can be evaluated as a function of $\mathbf{P}$ if* C-INFER *finds a mapping using the cftree-operators and G. Moreover, the process takes $O(n^2(n+m)p)$ time, where $n = |\mathbf{V}|$, m is the number of edges in G and $p = |\mathbf{P}|$ if there is at most one partially observed distribution in $\mathbf{P}$.*

7.5 Future Work

There are several problems in causal inference and data fusion that were not studied in this dissertation and could be incorporated into the framework. We provide a list with a short discussion about some of them.

Partial Observability Some of the tasks studied in this dissertation considered a distribution over a subset of the endogenous variables as input. Lee and Bareinboim [92] studied the task of *causal effect identifiability under partially-observed distributions* (*gid-PO*) consisting of identifying a marginal causal effect from an input consisting of several experimental and observational distributions over different subsets of the en-

dogenous variables. They developed a sound algorithm to solve *gid-PO* and further conjectured that deciding this task is an NP-complete problem. As the paper suggests, these developments have significant implications for the task of evaluating conditional queries in transportability settings. While solving the *s-TR* and *sbt-ID* tasks in this dissertation, we used our framework with the partially-observed distributions included in the input of these. Nevertheless, the completeness and complexity of the problem could be dealt with as the number of partially observed distributions was limited to one. An interesting question is how to extend this framework to deal with multiple partially-observed distributions as efficiently as possible.

Missing data Samples of the input distributions could be missing values for certain variables. The absence of these values could be completely at random or depend on other variables under study. This problem has been studied with graphical models to recover probabilistic quantities [93, 94, 95] and causal effects [96, 97]. Data with entries that are missing is another source of information that a causal inference engine could use to identify and generalize causal effects and probabilistic quantities.

Generalization of Counterfactual Queries The tasks studied in this dissertation considered queries that belong to the first and second layers of the Pearl Causal Hierarchy [39, 4]: associational and interventional input distributions and queries. However, there are many real world settings that evoke statements that are inherently counterfactuals, which belongs to the third layer of the causal hierarchy. Although there are some results on the identification of counterfactuals from observational and experimental data (e.g., [98, 99]), there is virtually no work on the generalizability of counterfactuals across domains, with arbitrary experimental conditions, and in the presence of selection bias. This is a significant direction for future investigation as human-like AI requires counterfactual reasoning capabilities [4].

References

[1] E. Bareinboim, A. Forney, and J. Pearl, "Bandits with unobserved confounders: A causal approach," in *Advances in Neural Information Processing Systems*, 2015, pp. 1342–1350.

[2] J. H. Chen and S. M. Asch, "Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations," *The New England journal of medicine*, vol. 376, no. 26, pp. 2507–2509, Jun. 2017.

[3] M. Buchanan, "The limits of machine prediction," *Nature Physics*, vol. 15, no. 4, p. 304, 2019.

[4] J. Pearl and D. Mackenzie, *The Book of Why*. New York: Basic Books, 2018, ISBN: 978-0-465-09760-9.

[5] B. Schölkopf, *Causality for Machine Learning*, 2019.

[6] R. A. Fisher, "Design of Experiments," *British Medical Journal*, vol. 1, no. 3923, p. 554, 1936.

[7] P. Spirtes, C. N. Glymour, and R Scheines, *Causation, Prediction, and Search*. New York: Springer-Verlag, 1993.

[8] D. Galles and J. Pearl, "Testing identifiability of causal effects," in *Uncertainty in Artificial Intelligence 11*, P Besnard and S Hanks, Eds., San Francisco: Morgan Kaufmann, 1995, pp. 185–195.

[9] J. Pearl and J. M. Robins, "Probabilistic evaluation of sequential plans from causal models with hidden variables," in *Uncertainty in Artificial Intelligence 11*, P Besnard and S Hanks, Eds., San Francisco: Morgan Kaufmann, 1995, pp. 444–453.

[10] J. Tian and J. Pearl, "A General Identification Condition for Causal Effects," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, Menlo Park, CA: AAAI Press/The MIT Press, 2002, pp. 567–573.

[11] Y. Huang and M. Valtorta, "Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm," in *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, Menlo Park, CA: AAAI Press, 2006, pp. 1149–1156.

[12] I. Shpitser and J. Pearl, "Identification of Joint Interventional Distributions in Recursive semi-Markovian Causal Models," in *Proceedings of the Twenty-First AAAI*

Conference on Artificial Intelligence, vol. 2, 2006, pp. 1219–1226, ISBN: 978-1-57735-281-5.

[13] E. Bareinboim and J. Pearl, "Causal Inference by Surrogate Experiments: z-Identifiability," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, N. d. F. Murphy and Kevin, Eds., AUAI Press, 2012, pp. 113–120, ISBN: 9780974903989.

[14] S. Lee, J. D. Correa, and E. Bareinboim, "General Identifiability with Arbitrary Surrogate Experiments," in *Proceedings of the Thirty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR: AUAI Press, 2019.

[15] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, "Invariant models for causal transfer learning," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1309–1342, 2018.

[16] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under Target and Conditional Shift," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, JMLR: W&CP volume 28, 2013.

[17] K. Zhang, M. Gong, and B. Scholkopf, "Multi-source Domain Adaptation: A Causal View," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15, AAAI Press, 2015, pp. 3150–3157, ISBN: 0-262-51129-0.

[18] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, "Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 846–10 856.

[19] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

[20] D. G. Altman, K. F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. C. Getzsche, and T. Lang, "The Revised CONSORT Statement for reporting randomized trials: Explanation and elaboration," *Annals of Internal Medicine*, 2001.

[21] J. B. Greenhouse, E. E. Kaizar, K. Kelleher, H. Seltman, and W. Gardner, "Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users," *Statistics in Medicine*, 2008.

[22] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[23] ——, "Transportability of causal effects: Completeness Results," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, Department of Computer Science, University of California, Los Angeles, CA, 2012.

[24] ——, "Controlling Selection Bias in Causal Inference," in *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, N. Lawrence and M. Girolami, Eds., La Palma, Canary Islands: JMLR, Apr. 2012, pp. 100–108.

[25] E. Bareinboim, S. Lee, V. Honavar, and J. Pearl, "Transportability from Multiple Environments with Limited Experiments," *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, 2013.

[26] E. Bareinboim, J. Tian, and J. Pearl, "Recovering from Selection Bias in Causal and Statistical Inference," in *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence*, C. E. Brodley and P. Stone, Eds., Palo Alto, CA: AAAI Press, 2014, pp. 2410–2416.

[27] E. Bareinboim and J. Tian, "Recovering Causal Effects from Selection Bias," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, S. Koenig and B. Bonet, Eds., 2015, pp. 3475–3481.

[28] J. Pearl and E. Bareinboim, "Transportability across studies: A formal approach," Cognitive Systems Laboratory, Department of Computer Science, UCLA, Tech. Rep. R-372, 2011.

[29] E. Bareinboim and J. Pearl, "A general algorithm for deciding transportability of experimental results," *Journal of Causal Inference*, vol. 1, no. 1, pp. 107–134, 2013.

[30] ——, "Transportability from Multiple Environments with Limited Experiments: Completeness Results," in *Advances in Neural Information Processing Systems 27*, Z Ghahramani, M Welling, C Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014, pp. 280–288.

[31] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd. New York, NY, USA: Cambridge University Press, 2000, ISBN: 978-0-521-89560-6.

[32] J. D. Correa and E. Bareinboim, "A Calculus For Stochastic Interventions: Causal Effect Identification and Surrogate Experiments," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, AAAI Press, 2020.

[33] ——, "General Transportability of Soft Interventions: Completeness Results," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[34] ——, "From Statistical Transportability to Estimating the Effects of Stochastic Interventions," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2019.

[35] ——, "Causal Effect Identification by Adjustment under Confounding and Selection Biases," in *Proceedings of the Thirty-First Conference on Artificial Intelligence*, Purdue AI Lab, Department of Computer Science, Purdue University, 2017, pp. 3740–3746.

[36] J. D. Correa, J. Tian, and E. Bareinboim, "Generalized Adjustment Under Confounding and Selection Biases," in *Proceedings of the 32th Conference on Artificial Intelligence*, 2018.

[37] ——, "Identification of Causal Effects in the Presence of Selection Bias," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, Hawai, USA: AAAI Press, 2019.

[38] ——, "Adjustment Criteria for Generalizing Experimental Findings," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, Long Beach, California, USA: PMLR, 2019, pp. 1361–1369.

[39] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, "On Pearl's Hierarchy and the Foundations of Causal Inference," Causal Artificial Intelligence Lab, Columbia University, Tech. Rep. R-60. In press. Jul. 2020.

[40] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[41] J. M. Robins, "A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect," *Mathematical Modeling*, vol. 7, pp. 1393–1512, 1986.

[42] A. P. Dawid, "Influence diagrams for causal modelling and inference," *International Statistical Review*, vol. 70, pp. 161–189, 2002.

[43] P Spirtes, C. N. Glymour, and R Scheines, *Causation, Prediction, and Search*, 2nd. Cambridge, MA: MIT Press, 2000.

[44] J. Pearl, "Causation, Action, and Counterfactuals," in *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Sixth Conference*, Y Shoham, Ed., San Francisco, CA: Morgan Kaufmann, 1996, pp. 51–73.

[45] ——, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[46] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd. MIT Press, 2001, ISBN: 9780262194402.

[47] J. Pearl, "Aspects of graphical models connected with causality," *Proceedings of the 49th Session of the International Statistical Institute*, vol. 1, no. August, pp. 399–401, 1993.

[48] J. Tian and J. Pearl, "On the Testable Implications of Causal Models with Hidden Variables," *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pp. 519–527, 2002.

[49] J. Pearl and E. Bareinboim, "External Validity and Transportability: A Formal Approach," in *Proceedings of the American Statistical Association, Joint Statistical Meetings*, Miami Beach, FL: {MIRA} Digital Publishing, 2011, pp. 157–171.

[50] S. Lee, J. D. Correa, and E. Bareinboim, "Generalized Transportability: Synthesis of Experiments from Heterogeneous Domains," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press, 2020.

[51] E. Bareinboim and J. Pearl, "Meta-Transportability of Causal Effects: A Formal Approach," in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013, pp. 135–143.

[52] J Woodward, *Making Things Happen*. New York, NY: Oxford University Press, 2003.

[53] J. J. Heckman, "The Scientific Model of Causality," *Sociological Methodology*, vol. 35, pp. 1–97, 2005.

[54] N Cartwright, *Hunting Causes and Using Them: {A}pproaches in Philosophy and Economics*. New York, NY: Cambridge University Press, 2007.

[55] V. Didelez, A. P. Dawid, and S. Geneletti, "Direct and indirect effects of sequential treatments," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 2006, pp. 138–146.

[56] J. Tian, "Identifying Dynamic Sequential Plans," in *In Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, Corvallis, Oregon: AUAI Press, 2008, 554–561, ISBN: 0-9749039-4-9.

[57] I. Shpitser and E. Sherman, "Identification of Personalized Effects Associated With Causal Pathways.," in *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018, pp. 530–539.

[58] F. Eberhardt and R. Scheines, "Interventions and causal inference," *Philosophy of science*, vol. 74, no. 5, pp. 981–995, 2007.

[59] A. P. Dawid, V. Didelez, and others, "Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview," *Statistics Surveys*, vol. 4, pp. 184–231, 2010.

[60] J. Pearl, "A probabilistic calculus of actions," in *Uncertainty in Artificial Intelligence 10*, R. L. de Mantaras and D.˜Poole, Eds., San Mateo, CA: Morgan Kaufmann, 1994, pp. 454–462.

[61] A. P. Dawid, "Statistical Causality from a Decision-Theoretic Perspective," *Annual Review of Statistics and Its Application*, vol. 2, no. 1, pp. 273–303, 2015.

[62] J. Pearl, "Comment: Graphical Models, Causality, and Intervention," *Statistical Science*, vol. 8, no. 3, pp. 266–269, 1993.

[63] A. V. Banerjee, S. Cole, E. Duflo, and L. Linden, "Remedying Education: Evidence from Two Randomized Experiments in India*," *The Quarterly Journal of Economics*, vol. 122, no. 3, pp. 1235–1264, 2007.

[64] E. Duflo, R. Glennerster, and M. Kremer, "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics*, T. P. Schultz and J. A. Strauss, Eds., vol. 4, Elsevier, 2007, ch. 61, pp. 3895–3962.

[65] M. Bertrand, D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman, "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment*," *The Quarterly Journal of Economics*, vol. 125, no. 1, pp. 263–306, Feb. 2010.

[66] D. Karlan and J. Zinman, "Expanding credit access: Using randomized supply decisions to estimate the impacts," *The Review of Financial Studies*, vol. 23, no. 1, pp. 433–464, 2010.

[67] J. Aldrich, "Autonomy," *Oxford Economic Papers*, vol. 41, pp. 15–34, 1989.

[68] J. Pearl and E. Bareinboim, "Transportability of Causal and Statistical Relations: A Formal Approach," in *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, Menlo Park, CA, Aug. 2011, pp. 247–254.

[69] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[70] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural networks*, vol. 6, no. 6, pp. 861–867, 1993.

[71] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6232–6240.

[72] E. Bareinboim, C. Brito, and J. Pearl, "Local Characterizations of Causal Bayesian Networks," in *Graph Structures for Knowledge Representation and Reasoning*, M. Croitoru, S. Rudolph, N. Wilson, J. Howse, and O. Corby, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–17, ISBN: 978-3-642-29449-5.

[73] G Cooper, "Causal Discovery from Data in the Presence of Selection Bias," in *Proceedings of the Workshop on Artificial Intelligence and Statistics*, 1995, pp. 140–150.

[74] C. Elkan, "The foundations of cost-sensitive learning," in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 17, 2001, pp. 973–978, ISBN: 1558608125,

[75] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Twenty-first international conference on Machine learning - ICML '04*, 2004, p. 114, ISBN: 1581138285.

[76] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, "Sample Selection Bias Correction Theory," in *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, 2008, pp. 38–53.

[77] A. Whittemore, "Collapsibility of multidimensional contingency tables," *Journal of the Royal Statistical Society. Series B*, vol. 40, no. 3, pp. 328–340, 1978.

[78] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New York, NY, USA: John Wiley & Sons, Inc., 1986, ISBN: 0-471-80254-9.

[79] L. D. Robinson and N. P. Jewell, "Some Surprising Results about Covariate Adjustment in Logistic Regression Models," *International Statistical Review / Revue Internationale de Statistique*, vol. 59, no. 2, pp. 227–240, 1991.

[80] M. Kuroki and Z. Cai, "On recovering a population covariance matrix in the presence of selection bias," *Biometrika*, vol. 93, no. 3, pp. 601–611, 2006.

[81] R. J. Evans and V. Didelez, "Recovering from Selection Bias Using Marginal Structure in Discrete Models," in *Proceedings of the UAI 2015 Conference on Advances in Causal Inference - Volume 1504*, ser. ACI'15, Aachen, Germany, Germany: CEUR-WS.org, 2015, pp. 46–55.

[82] M. Pirinen, P. Donnelly, and C. C. A. Spencer, "Including known covariates can reduce power to detect genetic effects in case-control studies," *Nature Genetics*, vol. 44, no. 8, pp. 848–851, 2012.

[83] J. Mefford and J. S. Witte, "The covariate's dilemma," *PLoS Genet*, vol. 8, no. 11, e1003096, 2012.

[84] J. J. Heckman, "Sample Selection Bias as a Specification Error," *Econometrica*, vol. 47, no. 1, p. 153, 1979.

[85] J. D. Angrist, "Conditional independence in sample selection models," *Economics Letters*, vol. 54, no. 2, pp. 103–112, 1997.

[86] J. M. Robins, "Data, Design, and Background Knowledge in Etiologic Inference," *Epidemiology*, vol. 12, no. 3, pp. 313–320, 2001.

[87] M. M. Glymour and S. Greenland, "Causal Diagrams," in *Modern Epidemiology*, K. R. Lash, S. Greenland, and T.L., Eds., 3rd, Philadelphia, PA: Lippincott Williams & Wilkins, 2008, pp. 183–209.

[88] J. Pearl and A. Paz, "Confounding Equivalence in Causal Equivalence," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, Corvallis, OR: AUAI, 2010, pp. 433–441.

[89] I. Shpitser, T. J. VanderWeele, and J. M. Robins, "On the validity of covariate adjustment for estimating causal effects," *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pp. 527–536, 2010.

[90] B. van der Zander, M. Liskiewicz, and J. Textor, "Constructing separators and adjustment sets in ancestral graphs," in *Proceedings of UAI 2014*, 2014, pp. 907–916, ISBN: 9780974903910.

[91] K. Takata, "Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph," *Discrete Applied Mathematics*, vol. 158, no. 15, pp. 1660–1667, 2010.

[92] S. Lee and E. Bareinboim, "Causal Effect Identifiability under Partial-Observability," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, PMLR, 2020.

[93] K. Mohan, J. Pearl, and J. Tian, "Graphical Models for Inference with Missing Data," in *Advances in Neural Information Processing System*, 2013, ISBN: 5660042538.

[94] I. Shpitser, K. Mohan, and J. Pearl, "Missing Data As a Causal and Probabilistic Problem," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, ser. UAI'15, Arlington, Virginia, United States: AUAI Press, 2015, pp. 802–811, ISBN: 978-0-9966431-0-8.

[95] J. Tian, "Recovering Probability Distributions from Missing Data," in *Proceedings of the Ninth Asian Conference on Machine Learning*, M.-L. Zhang and Y.-K. Noh, Eds., ser. Proceedings of Machine Learning Research, vol. 77, PMLR, 2017, pp. 574–589.

[96] K. Mohan and J. Pearl, "Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data," in *Advances in Neural Information Processing Systems*, 2014.

[97] M Saadati and J. Tian, "Adjustment Criteria for Recovering Causal Effects from Missing Data," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2019.

[98] J. Pearl, "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 2001, pp. 411–420.

[99] I. Shpitser and J. Pearl, "What Counterfactuals Can Be Tested," in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, Vancouver, BC, Canada: AUAI Press, 2007, pp. 352–359.

[100] D. Geiger, T. S. Verma, and J. Pearl, "d-Separation: From Theorems to Algorithms," in *Proceedings, 5th Workshop on Uncertainty in AI*, 1989, pp. 118–124, ISBN: 9780444887382.

[101] Y. Huang and M. Valtorta, "On the completeness of an identifiability algorithm for semi-Markovian models," *Annals of Mathematics and Artificial Intelligence*, vol. 54, no. 4, pp. 363–408, 2008.

[102] J. Tian, "Studies in Causal Reasoning and Learning," Ph.D. dissertation, Computer Science Department, University of California, Los Angeles, CA, Nov. 2002.

Appendix A: Background Results in Probability and Causality

A.1 The d-separation Criterion

A causal diagram G is a causal diagram for SCM inducing the observational distribution $P(\mathbf{V})$ [39]. Accordingly, G can be regarded as a set of *conditional independences* that $P(\mathbf{V})$ must satisfy. Graphically speaking, it is possible to read every such independence directly from G by means of the *d-separation* criterion [100].

Definition 28 (d-Separation)**.** A path p is blocked by a set of nodes $\mathbf{Z}$ if and only if

1. p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in $\mathbf{Z}$ (i.e., B is conditioned on), or
2. p contains a collider $A \to B \leftarrow C$ such that the collision node B is not in $\mathbf{Z}$, and no descendant of B is in $\mathbf{Z}$.

If $\mathbf{Z}$ blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on $\mathbf{Z}$, and thus are independent conditional on $\mathbf{Z}$, denoted as $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$.

A.2 do-calculus

Theorem 23. *[Do-Calculus] Let G be a causal diagram the set of all interventional distributions satisfies the do-Calculus rules according to G. Namely, for any disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ the following three rules hold:*

$$\textit{Rule 1} \quad P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}) \text{ in } G_{\overline{\mathbf{X}}}: \quad \text{(A.1)}$$

$$\textit{Rule 2} \quad P(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, \mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}) \text{ in } G_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}: \quad \text{(A.2)}$$

$$\textit{Rule 3} \quad P(\mathbf{y} \mid do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{w}) \qquad \text{if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}) \text{ in } G_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}; \quad \text{(A.3)}$$

where a graph $G_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$ *is obtained from G by removing the arrows incoming to* $\mathbf{X}$ *and outgoing from* $\mathbf{Z}$*, and* $\mathbf{Z}(\mathbf{W})$ *is the set of* $\mathbf{Z}$*-nodes non-ancestors of* $\mathbf{W}$ *in the corresponding graph.*

Appendix B: Cftree Operators

B.1 Relevant Results from the Literature

The following lemma from [48] shows variables with no descendants in the c-factor's subgraph can be summed out.

Lemma 10 (C-factor marginalization)**.** *Let* $\mathbf{W} \subseteq \mathbf{C} \subseteq \mathbf{V}$ *be an ancestral set in* $G[\mathbf{C}]$*, that is,* $\mathbf{W} = An(\mathbf{W})_{G[\mathbf{C}]}$*, then:*

$$Q[\mathbf{W}] = \sum_{\mathbf{C} \setminus \mathbf{W}} Q[\mathbf{C}]. \tag{B.1}$$

Lemma 11 (C-component factorization)**.** *Let* $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_l\} \subseteq \mathbf{V}$ *and let* $\mathbf{C}_1, \ldots, \mathbf{C}_l$ *the c-components of* $G[\mathbf{C}]$*. Then*

$$Q[\mathbf{C}] = \prod_j Q[\mathbf{C}_j], \tag{B.2}$$

and for any topological order $C_1 < C_2 < C_n$ *of the variables in* $\mathbf{C}$

$$Q[\mathbf{C}_j] = \prod_{\{C_i \in \mathbf{C}_j\}} \frac{Q[C_1, \ldots, C_i]}{Q[C_1, \ldots, C_{i-1}]}, \text{ where } Q[C_1, \ldots, C_i] = \sum_{C_{i+1}, \ldots, C_n} Q[\mathbf{C}]. \tag{B.3}$$

While lemma 10 gives some notion of marginalization to c-factors, lemma 11 generalize the natural factorization of Markovian models to the non-Markovian case.

B.2 and Operators

Lemma 1 (and operators)**.** *Let endogenous sets* $\mathbf{T}; \mathbf{C}$ *be such that* $\mathbf{C}$ $\mathbf{T}$ *and let* $\mathbf{C}^0 = \mathbf{T} n \mathbf{C}$*. Then, for any cftree* T *with a node* $Q[\mathbf{T}]$*:*

-operator (marginalization): If there is no directed arrow with tail in $\mathbf{C}^0$ *and head in* $\mathbf{C}$ *in the diagram* $G[\mathbf{T}]$*,* $Q[\mathbf{T}]$! $Q[\mathbf{C}]$ *is a valid edge for* T *and it is associated with the function*

$$Q[\mathbf{C}] = \sum_{\mathbf{c}^0} Q[\mathbf{T}]: \tag{3.20}$$

-operator (independence): If there is no bidirected arrow connecting $\mathbf{C}$ *and* $\mathbf{C}^0$ *in the diagram* $G[\mathbf{T}]$*,* $Q[\mathbf{T}]$! $Q[\mathbf{C}]$ *is a valid edge for* T*. The associated function is defined for any topological order* $T_1 < T_2 <$ $< T_k$ *on* $G[\mathbf{T}]$ *as*

$$Q[\mathbf{C}] = \prod_{T_i 2 \mathbf{C}} \frac{\sum_{t_{i+1},:::,t_k} Q[\mathbf{T}]}{\sum_{t_i,:::,t_k} Q[\mathbf{T}]}: \tag{3.21}$$

Furthermore, let $\mathbf{C}_1; : : : ; \mathbf{C}_k$ *be the c-components of* $G[\mathbf{T}]$*, then the set of edges* $fQ[\mathbf{T}]$ $Q[\mathbf{C}_i]g_{i=1,:::,k}$ *are valid for* T *and correspond to the mapping*

$$Q[\mathbf{T}] = \prod_{i=1}^{k} Q[\mathbf{C}_i]: \tag{3.22}$$

Proof. First notice the condition for the -operator implies $\mathbf{C}$ is an ancestral set in $G[\mathbf{T}]$. To see why, assume for the sake of contradiction, there exists a path T_0 ! T_1 ! ! T_k for $T_0; T_1;$ $; T_{k\ 1}$ 2 $An(T_k)_{G[\mathbf{T}]}, k$ 1 for some T_k 2 $\mathbf{C}$ and T_0 2 $\mathbf{C}^0$. At least one edge $(A_i$! $A_{i+1})$ with A_i 2 $\mathbf{C}^0$ and A_{i+1} 2 $\mathbf{C}$ which witnesses a contradiction to this condition. Since $\mathbf{C}$ is an ancestral set, eq. (3.20) follows from lemma 10.

For the -operator, the absence of bidirected arrows between $\mathbf{C}$ and $\mathbf{C}^0$ in $G[\mathbf{T}]$ implies each c-component of $G[\mathbf{T}]$ either belongs in $\mathbf{C}$ or $\mathbf{C}^0$. Let $\mathbf{T}_1; : : : ; \mathbf{T}_k$ be all c-components in

C, then

$$Q[\mathbf{C}] = Q[\mathbf{T}_1] \quad Q[\mathbf{T}_k] = \prod_{j=1}^{k} \prod_{T_i \subseteq \mathbf{T}_j} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T}]} = \prod_{T_i \subseteq \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T}]}. \tag{B.4}$$

The first equality holds by lemma 11, eq. (B.2) as $\mathbf{T}_j$, $j = 1,\ldots,k$ are also the c-components of $G[\mathbf{C}]$. The second equality follows lemma 11 and eq. (B.3) because each $\mathbf{T}_j$ is computable from $Q[\mathbf{T}]$. Equation (3.21) follows after writing the products together. □

B.3 and Operators

Lemma 4 (operator). *Let* $\mathbf{T} \subseteq \mathbf{V}$ *be an endogenous set of variables. Then, for any cftree* T *with a node* $Q[\mathbf{T}, \mathbf{x}]$*:*

-operator (regime invariance): If $\mathbf{T} \setminus \mathbf{X} = \emptyset$, $Q[\mathbf{T}, \mathbf{x}] \neq Q[\mathbf{T}, \hat{\mathbf{x}}]$ *and* $Q[\mathbf{T}, \mathbf{x}]$ $Q[\mathbf{T}, \hat{\mathbf{x}}]$ *are valid edges for* T *and mapped as*

$$Q[\mathbf{T}, \hat{\mathbf{x}}] = Q[\mathbf{T}, \mathbf{x}]. \tag{4.72}$$

Proof. By definition of c-factor,

$$Q[\mathbf{T}, \mathbf{x}](\mathbf{v}) = \sum_{\mathbf{u}(\mathbf{T}) \, \{i \mid V_i \in \mathbf{T}\}} \prod P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \mathbf{x}) P(u(\mathbf{T}), \mathbf{x}). \tag{B.5}$$

Since no variable in $\mathbf{T}$ is also in $\mathbf{X}$, every term $P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \mathbf{x}) = P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \hat{\mathbf{x}})$ and $P(\mathbf{u}(\mathbf{T}), \mathbf{x}) = P(\mathbf{u}(\mathbf{T}), \hat{\mathbf{x}}) = P(\mathbf{u}(\mathbf{T}))$ as neither $\mathbf{x}$ nor $\hat{\mathbf{x}}$ affect f_i or the distribution of variables in $\mathbf{U}$. Replacing those terms, we obtain exactly the definition of $Q[\mathbf{T}, \hat{\mathbf{x}}](\mathbf{v})$. □

Lemma 3 (operator). *Let* $\mathbf{T} \subseteq \mathbf{V}$ *be an endogenous set of variables. Then, for any cftree* T *with a node* $Q^a[\mathbf{T}]$*:*

-operator (domain invariance): If $\mathbf{T} \setminus {}^{a,b} = \emptyset$, $Q^a[\mathbf{T}]$ $Q^b[\mathbf{T}]$ *and* $Q^a[\mathbf{T}] \neq Q^b[\mathbf{T}]$ *are*

valid edges for T and corresponds to the mapping

$$Q^a[\mathbf{T}] = Q^b[\mathbf{T}]: \tag{3.57}$$

Proof. From the definition of c-factor we have

$$Q^j[\mathbf{T}](\mathbf{v}) = \sum_{\mathbf{u}(\mathbf{T})\ \{ij V_i 2 \mathbf{T}\}} \prod P^j(v_i \mid \mathbf{pa}_i; \mathbf{u}_i) P^{(j)}(\mathbf{u}(\mathbf{T})); \quad j = a; b: \tag{B.6}$$

$V_i \in \mathbf{T}_{a;b}$ only if $f_i^{(k)} \neq f_i^{(l)}$ or if $P^{(k)}(\mathbf{U}_i) \neq P^{(l)}(\mathbf{U}_i)$. Thus, $\mathbf{T} \setminus \mathbf{T}_{a;b} = \emptyset$ implies $P^k(v_i \mid \mathbf{pa}_i; \mathbf{u}_i) = P^l(v_i \mid \mathbf{pa}_i; \mathbf{u}_i)$ and $P^k(\mathbf{U}_i) = P^l(\mathbf{U}_i)$. Then, every term in eq. (B.6) is the same in both domains and the claim follows. □

B.4 Operator

Lemma 5 (-operator). *Let* $\mathbf{T}; \mathbf{L} \subseteq \mathbf{V}$ *be disjoint sets,* $\mathbf{C} \subseteq \mathbf{T}$, $\mathbf{C}' = \mathbf{T} \cap \mathbf{C}$, $\mathbf{H} \subseteq \mathbf{L}$ *and* $\mathbf{H}' = \mathbf{L} \cap \mathbf{H}$. *Then, for any cftree T :*

-operator (unconditioning): Let $\mathbf{T}_1; \mathbf{T}_2 \subseteq \mathbf{T}$ *be disjoint sets such that* $\mathbf{T}_1 \cup \mathbf{T}_2 = \mathbf{T}$. *If* $\mathbf{T}_1 \subseteq \mathbf{C} \subseteq \mathbf{T}$ *and there is no path between* $\mathbf{C} \cup \mathbf{H}$ *and* $\mathbf{C}' \cup \mathbf{H}'$ *in* $G[\mathbf{T}; \mathbf{L}]_{\underline{\mathbf{T}_2}}$, $Q[\mathbf{T}_1 \mid \mathbf{T}_2 \cup \mathbf{L}] = Q[\mathbf{C} \cup \mathbf{H}]$ *is a valid edge for T . The associated function is*

$$Q[\mathbf{T}_1 \mid \mathbf{T}_2 \cup \mathbf{L}] = \frac{Q[\mathbf{C} \cup \mathbf{H}]}{\sum_{\mathbf{t}_1} Q[\mathbf{C} \cup \mathbf{H}]} \tag{5.38}$$

Proof. By definition

$$Q[\mathbf{T}_1 \mid \mathbf{T}_2 \cup \mathbf{L}] = \frac{Q[\mathbf{T}_1; \mathbf{T}_2 \cup \mathbf{L}]}{\sum_{\mathbf{t}_1} Q[\mathbf{T}_1; \mathbf{T}_2 \cup \mathbf{L}]}: \tag{B.7}$$

From the condition we have that the sets $\mathbf{C} \cup \mathbf{H}$ and $\mathbf{C}' \cup \mathbf{H}'$ are in different marginalized

c-components (when summing $\mathbf{T}_1$), then by the -operator we have

$$Q[\mathbf{T}_1 \, j \, \mathbf{T}_2 \, k \, \mathbf{L}] = Q[\mathbf{C}^{\prime} \, k \, \mathbf{H}^{\prime}] Q[\mathbf{C} \, k \, \mathbf{H}]: \tag{B.8}$$

Moreover,

$$\sum_{\mathbf{t}_1} Q[\mathbf{T}_1 \, j \, \mathbf{T}_2 \, k \, \mathbf{L}] = Q[\mathbf{C}^{\prime} \, k \, \mathbf{H}^{\prime}] \sum_{\mathbf{t}_1} Q[\mathbf{C} \, k \, \mathbf{H}]; \tag{B.9}$$

because no variable in $\mathbf{C}^{\prime}$ is also in $\mathbf{T}_1$ or has a parent in $\mathbf{T}_1$, otherwise $\mathbf{C}^{\prime}$ would not be marginalized c-component independent of $\mathbf{C}$. Then $Q[\mathbf{C}^{\prime} \, k \, \mathbf{H}^{\prime}]$ is not a function of $\mathbf{T}_1$. Finally,

$$Q[\mathbf{T}_1 \, j \, \mathbf{T}_2 \, k \, \mathbf{L}] = \frac{Q[\mathbf{C}^{\prime} \, k \, \mathbf{H}^{\prime}] \sum_{\mathbf{t}_1} Q[\mathbf{C} \, k \, \mathbf{H}]}{Q[\mathbf{C}^{\prime} \, k \, \mathbf{H}^{\prime}]} \tag{B.10}$$

$$= \sum_{\mathbf{t}_1} Q[\mathbf{C} \, k \, \mathbf{H}]; \tag{B.11}$$

which proves the -operator. □

B.5 , $^{\prime}$, and $^{\prime}$ Operators

Lemma 8 (-, $^{\prime}$-, and -operator for selection-biased c-factors)**.** *Let* $\mathbf{T}$ $\mathbf{V}$, $\mathbf{C}$ $\mathbf{T}$, $\mathbf{C}^{\prime} = \mathbf{T} \cap \mathbf{C}$. *Let* $T_1 < T_2 <$ $< T_k$ *be a topological order of* $G[\mathbf{T}]$ *where* $An(S) < \mathbf{T} \cap An(S)$. *Then, for any cftree T :*

-operator (marginalization): If there is no directed arrow with tail in $\mathbf{C}^{\prime}$ *and head in* $\mathbf{C} \cup \{S\}$ *in* $G[\mathbf{T} \cup \{S\}]$, $Q[\mathbf{T} \, j \, S]$! $Q[\mathbf{C} \, j \, S]$ *is a valid edge for T with mapping*

$$Q[\mathbf{C} \, j \, S] = \sum_{\mathbf{c}^{\prime}} Q[\mathbf{T} \, j \, S]: \tag{6.35}$$

-operator (recovery): If $\mathbf{C} \setminus An(S) =$; *and there is no bidirected edge between* $\mathbf{C}$ *and* $\mathbf{C}^{\prime} \cup \{S\}$

in $G[\mathbf{T} \cup \{S\}]$, $Q[\mathbf{T} \mid S] \to Q[\mathbf{C}]$ *is a valid edge for* T *and the corresponding mapping is*

$$Q[\mathbf{C}] = \prod_{T_i \in \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \mid S]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \mid S]}. \tag{6.36}$$

$'$-operator *(recovery residue): If* $\mathbf{C}' \setminus An(S) = \emptyset$ *and there is no bidirected edge between* $\mathbf{C} \cup \{S\}$ *and* $\mathbf{C}'$ *in* $G[\mathbf{T} \cup \{S\}]$, $Q[\mathbf{T} \mid S] \to Q[\mathbf{C} \mid S]$ *is a valid edge for* T *the corresponding mapping is*

$$Q[\mathbf{C} \mid S] = Q[\mathbf{T} \mid S] \prod_{T_i \in \mathbf{C}'} \frac{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \mid S]}{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \mid S]}. \tag{6.37}$$

Proof. First, note that

$$Q[\mathbf{T} \mid S] = Q[\mathbf{T}, S] / P(S = 1). \tag{B.12}$$

-operator. The condition for the -operator implies that $\mathbf{C}'$ contains no ancestor of $\mathbf{C} \cup \{S\}$, then by the usual -operator we have

$$Q[\mathbf{C}, S] = \sum_{\mathbf{c}'} Q[\mathbf{T}, S], \tag{B.13}$$

$$\frac{Q[\mathbf{C}, S]}{P(S = 1)} = \sum_{\mathbf{c}'} \frac{Q[\mathbf{T}, S]}{P(S = 1)} \tag{B.14}$$

$$Q[\mathbf{C} \mid S] = \sum_{\mathbf{c}'} Q[\mathbf{T} \mid S], \tag{B.15}$$

which proves the validity of the operator.

and $'$ operators. By the -operator (or lemma 11), $Q[\mathbf{T}, S] = Q[\mathbf{C}]Q[\mathbf{C}', S]$ because there are no bidirected arrows between $\mathbf{C}$ and $\mathbf{C}' \cup \{S\}$ in $G[\mathbf{T} \cup \{S\}]$.

Moreover, as $\mathbf{C} \setminus An(S) = \emptyset$ in $G[\mathbf{T} \cup \{S\}]$, there exists a topological order $T_1 < T_2 <$ where $An(S) < \mathbf{T} \setminus An(S)$. Then, for any $T_i \in \mathbf{C}$ every variable in $\mathbf{C}'$ and S come before in

the order. Then, each factor in eq. (6.36) is equal to

$$\frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T} \mid S]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T} \mid S]} \tag{B.16}$$

$$= \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{C}^0; S] Q[\mathbf{C}] / P(S=1)}{\sum_{t_i,\ldots,t_k} Q[\mathbf{C}^0; S] Q[\mathbf{C}] / P(S=1)} \qquad \text{(def. of conditional c-factor)} \tag{B.17}$$

$$= \frac{Q[\mathbf{C}^0; S]}{Q[\mathbf{C}^0; S]} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{C}] / P(S=1)}{\sum_{t_i,\ldots,t_k} Q[\mathbf{C}] / P(S=1)} \qquad (\{T_i; T_{i+1}, \ldots\} \setminus \mathbf{C}^0 \cup \{S\} = \emptyset) \tag{B.18}$$

$$= \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{C}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{C}]} \qquad \text{(cancel out common factors)} \tag{B.19}$$

$$= \frac{Q[\mathbf{C}^0]}{Q[\mathbf{C}^0]} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{C}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{C}]} \qquad \text{(multiply by } Q[\mathbf{C}^0]/Q[\mathbf{C}^0]) \tag{B.20}$$

$$= \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T}]}. \qquad (\ \text{-operator}) \tag{B.21}$$

Then, the product over all $\mathbf{T}_i \in \mathbf{C}$ corresponds exactly to the function of the -operator so that

$$\prod_{T_i \in \mathbf{C}} \frac{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T}]}{\sum_{t_i,\ldots,t_k} Q[\mathbf{T}]} = Q[\mathbf{T}], \tag{B.22}$$

which proves the -operator.

For the ^{0}-operator, switch the notation for $\mathbf{C}$ and $\mathbf{C}^0$ and notice that from eq. (B.12)

$$Q[\mathbf{T} \mid S] = Q[\mathbf{C}; S] Q[\mathbf{C}^0] / P(S=1). \tag{B.23}$$

By multiplying both sides of this equation by the inverse of of eq. (6.36) with $\mathbf{C} = \mathbf{C}^0$ we get

$$Q[\mathbf{T} \mid S] \prod_{T_i \in \mathbf{C}^0} \frac{\sum_{t_i,\ldots,t_k} Q[\mathbf{T}]}{\sum_{t_{i+1},\ldots,t_k} Q[\mathbf{T}]} = \frac{Q[\mathbf{C}; S] Q[\mathbf{C}^0] / P(S=1)}{Q[\mathbf{C}^0]} = Q[\mathbf{C}; S] / P(S=1) \tag{B.24}$$

$$= Q[\mathbf{C} \mid S]; \tag{B.25}$$

which proves the ^{0}-operator. □

Lemma 9 (- and ^{0}-operators for combining c-factors)**.** *Let endogenous sets* $\mathbf{T}; \mathbf{C}; \mathbf{C}^{0}; \mathbf{Z}; \mathbf{L}$ *be such that* $\mathbf{C} \quad \mathbf{T}$, $\mathbf{C}^{0} = \mathbf{T} \cap \mathbf{C}$ *and* $\mathbf{L} = An(\mathbf{Z})_{G[\mathbf{T} \cup fSg]}$. *Also, let* $\mathbf{A} = An(\mathbf{C}; S) \cap fSg$ *in* $G[\mathbf{T} \cup fSg]$ *and* $\mathbf{R} = An(S) \cap fSg$ *in* $G[\mathbf{T} \cup fSg]_{\underline{\mathbf{Z} \setminus \mathbf{T}}}$, $C_1 < C_2 < \cdots$ *be a topological order of* $G[\mathbf{A} \cap \mathbf{R}; \mathbf{Z} \setminus \mathbf{A}]$ *and*

$$Q^{0} = \frac{\sum_{\mathbf{t} \cap \mathbf{a}} Q[\mathbf{T} \mid S] \sum_{\mathbf{z} \cap \mathbf{r}} Q[\mathbf{Z} \mid \mathbf{L}]}{\sum_{\mathbf{t} \cap \mathbf{r}} Q[\mathbf{T} \mid S]}. \tag{6.50}$$

Then, for any pair $Q[\mathbf{T} \mid S]$ *and* $Q[\mathbf{Z} \mid \mathbf{L}]$*:*

- *-operator (unbiasing): If there is no path (regardless of the direction) between* $\mathbf{C}$ *and* S *in* $G[\mathbf{A} \cup fSg]_{\underline{\mathbf{Z} \setminus \mathbf{A}}}$ *and there is no bidirected edge between* $\mathbf{C}$ *and* $\mathbf{C}^{0} \cup fSg$ *in* $G[\mathbf{T} \cup fSg]_{\underline{\mathbf{Z} \setminus \mathbf{T}}}$, *then* $fQ[\mathbf{T} \mid S]; Q[\mathbf{Z} \mid \mathbf{L}]g \quad ! \quad Q[\mathbf{C}]$ *are valid edges for a cftree* T *and the corresponding mapping is*

$$Q[\mathbf{C}] = \prod_{C_i \in \mathbf{C}} \frac{\sum_{c_i+1, \ldots, c_k} Q^{0}}{\sum_{c_i, \ldots, c_k} Q^{0}}. \tag{6.51}$$

- ^{0}*-operator (unbiasing residue): If* $\mathbf{C}^{0} \setminus An(S) = ;$, *there is no bidirected edge between* $\mathbf{C}^{0}$ *and* $\mathbf{C} \cup fSg$, *and there is no bidirected edge between* $\mathbf{Z} \setminus \mathbf{R}$ *and* $(\mathbf{R} \cup fSg) \cap \mathbf{Z}$ *in* $G[\mathbf{T} \cup fSg]_{\underline{\mathbf{Z} \setminus \mathbf{T}}}$, *then* $fQ[\mathbf{T} \mid S]; Q[\mathbf{Z} \mid \mathbf{L}]g \quad ! \quad Q[\mathbf{C} \mid S]$ *are valid edges for a cftree* T *and the corresponding mapping is*

$$Q[\mathbf{C} \mid S] = Q[\mathbf{T} \mid S] \prod_{C_i \in \mathbf{C}^{0}} \frac{\sum_{c_i, \ldots, c_k} Q^{0}}{\sum_{c_i+1, \ldots, c_k} Q^{0}}. \tag{6.52}$$

Proof. First, let us consider the expression Q^{0}. Since $\mathbf{A}$ is an ancestral set

$$\sum_{\mathbf{t} \cap \mathbf{a}} Q[\mathbf{T} \mid S] = Q[\mathbf{A} \mid S] \tag{B.26}$$

$$= Q[\mathbf{A} \cap (\mathbf{R} \cap \mathbf{Z})] Q[\mathbf{R} \cap \mathbf{Z}; S] / P(S = 1); \tag{B.27}$$

where the last decomposition follows from the fact that there are no bidirected arrows

between $\mathbf{A} \cap (\mathbf{R} \cap \mathbf{Z})$ and $(\mathbf{R} \cap \mathbf{Z}) \cup \{S\}$, otherwise there is a path from $\mathbf{C}$ to S in $G[\mathbf{A} \cup \{S\}]_{\underline{\mathbf{Z} \setminus \mathbf{A}}}$. This is because every $\mathbf{A} \cap (\mathbf{R} \cap \mathbf{Z})$ has a directed path to $\mathbf{C}$ or is in $\mathbf{C}$ in that graph, and every $\mathbf{R} \cap \mathbf{Z}$ has a directed path to S in the same graph. Appending the directed paths with any such bidirected witnesses a contradiction to the condition in the operator.

Moreover, using the -operator, we can write

$$\sum_{\mathbf{z} \cap \mathbf{r}} Q[\mathbf{Z} \mid \mathbf{L}] = Q[\mathbf{Z} \setminus \mathbf{R} \mid An(\mathbf{Z} \setminus \mathbf{R})] \tag{B.28}$$

$$\sum_{\mathbf{t} \cap \mathbf{r}} Q[\mathbf{T} \mid S] = \sum_{\mathbf{a} \cap \mathbf{r}} Q[\mathbf{A} \mid S] \tag{B.29}$$

$$= \sum_{\mathbf{a} \cap \mathbf{r}} Q[\mathbf{A}; S]/P(S = 1) \tag{B.30}$$

$$= Q[\mathbf{R}; S \mid \mathbf{A} \cap \mathbf{R}]/P(S = 1) \tag{B.31}$$

$$= Q[\mathbf{Z} \setminus \mathbf{R} \mid An(\mathbf{Z} \setminus \mathbf{R})]Q[\mathbf{R} \cap \mathbf{Z}; S]/P(S = 1); \tag{B.32}$$

where the last equality holds by the -operator since there are no bidirected arrows between $\mathbf{Z}$ and $\mathbf{R} \cap \mathbf{Z}$. The lack of these bidirected arrows is due to the fact that $\mathbf{Z}$ is a descendant of some variable in $\mathbf{C}$ and every $\mathbf{R}$ is an ancestor of S. Using such bidirected arrow we can form a path from $\mathbf{C}$ to S that contradicts the condition of the -operator. Then

$$Q^{\emptyset} = \frac{(Q[\mathbf{A} \cap (\mathbf{R} \cap \mathbf{Z})]Q[\mathbf{R} \cap \mathbf{Z}; S]/P(S = 1))Q[\mathbf{Z} \setminus \mathbf{R} \mid An(\mathbf{Z} \setminus \mathbf{R})]}{Q[\mathbf{Z} \setminus \mathbf{R} \mid An(\mathbf{Z} \setminus \mathbf{R})]Q[\mathbf{R} \cap \mathbf{Z}; S]/P(S = 1)} = Q[\mathbf{A} \cap (\mathbf{R} \cap \mathbf{Z})]; \tag{B.33}$$

Let $\mathbf{A}^{\emptyset} = \mathbf{A} \cap (\mathbf{R} \cap \mathbf{Z})$, then $Q^{\emptyset} = Q[\mathbf{A}^{\emptyset}]$. Notice that, $\mathbf{A}^{\emptyset} \cap \mathbf{C}$ $\mathbf{C}^{\emptyset}$, and since there are no bidirected arrows between $\mathbf{C}$ and $\mathbf{C}^{\emptyset}$, the -operator licenses eq. (6.51).

For the ^{0}-operator, switch the notation for $\mathbf{C}$ and $\mathbf{C}^{0}$ and notice that from eq. (B.12)

$$Q[\mathbf{T} \; j \; S] = Q[\mathbf{C}; S]Q[\mathbf{C}^{0}]=P(S = 1): \tag{B.34}$$

By multiplying both sides of this equation by the inverse of of eq. (6.51) with $\mathbf{C} = \mathbf{C}^{0}$ we get

$$Q[\mathbf{T} \; j \; S] \prod_{C_i 2 \mathbf{C}^{0}} \frac{\sum_{c_i;\ldots;c_k} Q[\mathbf{T}]}{\sum_{c_{i+1};\ldots;c_k} Q[\mathbf{T}]} = \frac{Q[\mathbf{C}; S]Q[\mathbf{C}^{0}]=P(S = 1)}{Q[\mathbf{C}^{0}]} = Q[\mathbf{C}; S]=P(S = 1) \tag{B.35}$$

$$= Q[\mathbf{C} \; j \; S]; \tag{B.36}$$

which proves the ^{0}-operator. □

Appendix C: Soundness and Completeness Results for the Tasks

C.1 Tasks of identification or transportability of causal effects

First, note that the -*TR* subsumes all of the *obs-ID*, *g-ID*, and *g-TR* tasks. Therefore, in this section we will prove the soundness and completeness of *C-Infer* for -*TR* which implies the others.

First, we prove a result that relates the transportability of a c-factor $Q[\mathbf{A}]$ and a sum over it.

Lemma 12. *Suppose* $Q[\mathbf{A}; \mathbf{x}]$ *is not identifiable from a set of available distributions in a causal diagram* G. *Let* $A_1; A_2 \in \mathbf{A}$ *such that there exists an edge* $A_1 \to A_2$ *in* G. *Then* $\sum_{a_1} Q[\mathbf{A}]$ *is not identifiable from the same input either.*

Proof. Let M_1 and M_2 be the two models witnessing the non-identifiability of $Q[\mathbf{A}; \mathbf{x}]$. They must agree on available distributions, but for some value-assignment $\mathbf{v}^0$, we have $Q^1[\mathbf{A}; \mathbf{x}](\mathbf{v}^0) = \quad , Q^2[\mathbf{A}; \mathbf{x}](\mathbf{v}^0) = \quad$ with $\quad \neq \quad$. Assume, without loss of generality that $\quad > \quad$. We will extend a strategy used by [101] to construct two models M_1^0 and M_2^0 where the domain of A_2 is $\text{Val}(A_2) \quad \{0; 1\}$, where $\text{Val}(A_2)$ is the domain of A_2 in $M_1; M_2$. Let $F(A_1)$ be a probability function from $\text{Val}(A_1)$ to $\{0; 1\}$, such that $P(F(a_1) = i) > 0; i = 0; 1$ and $P(F(a_1) = 0) = 1 \quad P(F(a_1) = 1)$. In $M_i^0; i = 1; 2$ we define:

$$P_i^{M_i^0}((a_2; k) \mid \mathbf{pa}_{a_2}; u_{a_2}) = P^{M_i}(a_2 \mid \mathbf{pa}_{a_2}; u_{a_2}) P(F(a_1) = k). \tag{C.1}$$

And for $V_j \in \mathbf{V} \setminus \{A_2\}$ let $P^{M_i^0}(v_j \mid \mathbf{pa}_j; u_j) = P^{M_i}(v_j \mid \mathbf{pa}_j; u_j)$. We can verify that for any

$\mathbf{z}_j \; 2 \; \mathbf{Z}$

$$P^{M_1^{\prime}}(\mathbf{v} \; n \; a_2; (a_2; k); \; \mathbf{z}_j)$$

$$= Q^{M_1^{\prime}}[\mathbf{V} \; n \; fA_2g; (A_2; K); \; \mathbf{z}_j](\mathbf{v} \; n \; a_2; (a_2; k)) \tag{C.2}$$

$$= Q^{M_1}[\mathbf{V} \; n \; fA_2g; (A_2; K); \; \mathbf{z}_j](\mathbf{v})P(F(a_1) = k) \tag{C.3}$$

$$= Q^{M_2}[\mathbf{V} \; n \; fA_2g; (A_2; K); \; \mathbf{z}_j](\mathbf{v})P(F(a_1) = k) \tag{C.4}$$

$$= Q^{M_2^{\prime}}[\mathbf{V} \; n \; fA_2g; (A_2; K); \; \mathbf{z}_j](\mathbf{v} \; n \; a_2; (a_2; k)) \tag{C.5}$$

$$= P^{M_2^{\prime}}(\mathbf{v} \; n \; a_2; (a_2; k); \; \mathbf{z}_j): \tag{C.6}$$

Consider the assignment $\mathbf{v}^{\prime} \; n \; fa_2g; (a_2^{\prime}; 0)$, by construction we have

$$Q^{M_i^{\prime}}[\mathbf{A}; \; \mathbf{x}](\mathbf{v}^{\prime} \; n \; fa_2g; (a_2^{\prime}; 0))$$

$$= Q^{M_i}[\mathbf{A}; \; \mathbf{x}](\mathbf{v}^{\prime})P(F(a_1^{\prime}) = 0): \tag{C.7}$$

Let $P(F(a_1^{\prime}) = 0) = 1=2$ and $P(F(a_1) = 0) = (\quad)=4$, for $a_1 \; 6= \; a_1^{\prime}$. It yields:

$$\sum_{a_1} Q^{M_i^{\prime}}[\mathbf{A}; \; \mathbf{x}](\mathbf{v}^{\prime} \; n \; fa_1; a_2g; (a_2^{\prime}; 0))$$

$$= \sum_{a_1} Q^{M_i}[\mathbf{A}; \; \mathbf{x}](\mathbf{v}^{\prime} \; n \; fa_1; a_2g)P(F(a_1) = 0) \tag{C.8}$$

For $M_1^{\prime}$ this means

$$\sum_{a_1} Q^{M_1^{\prime}}[\mathbf{A}; \; \mathbf{x}](\mathbf{v}^{\prime} \; n \; fa_1; a_2g; (a_2^{\prime}; 0))$$

$$= \frac{1}{2} \; + \; \frac{}{4} \sum_{a_1 6= a_1^{\prime}} Q^{M_i}[\mathbf{A}; \; \mathbf{x}](\mathbf{v}^{\prime} \; n \; fa_1; a_2g) \tag{C.9}$$

$$> \frac{1}{2} \tag{C.10}$$

As for $M_2^{\emptyset}$:

$$\sum_{a_1} Q^{M_1^{\emptyset}}[\mathbf{A};\ \mathbf{x}](\mathbf{v}^{\emptyset} \cap \{a_1; a_2\}; (a_2^{\emptyset}; 0))$$

$$= \frac{1}{2}b + \frac{a\ b}{4} \sum_{a_1 \neq a_1^{\emptyset}} Q^{M_i}[\mathbf{A};\ \mathbf{x}](\mathbf{v}^{\emptyset} \cap \{a_1; a_2\}) \tag{C.11}$$

$$< \frac{1}{2} + \frac{}{4} \tag{C.12}$$

$$< \frac{1}{2} : \tag{C.13}$$

Then, $M_1^{\emptyset}$ and $M_2^{\emptyset}$ are compatible with G, match in the available distributions and yield different $\sum_{a_1} Q[\mathbf{A};\ \mathbf{x}]$. □

Theorem 3 (-, -, -, and -operators soundness and completeness for g-TR)**.** *Given a causal inference task with signature $I_{g\text{-}TR}$, the query is transportable from* P *and G^{Δ} if and only if* C-INFER *finds a mapping using the , , , and operators. Moreover, the task is decided in $O(n^2(n+m)p)$ time, where $n = |\mathbf{V}|$, m is the number of edges in G, and $p = |\mathbf{P}|$.*

Proof. **Soundness.** The soundness follows from the validity of the , , and cftree operators.

Completeness. *C-Infers* fails at a *-TR* task if there exists some c-factor $Q[\mathbf{A}_i;\ \mathbf{x}]$ in the q-tree that could not be mapped from any of the d-trees. This is the case when $\mathbf{A}_i$ is a c-component of $G_{\mathbf{x}[\mathbf{D}]}$, $\mathbf{D} = An(\mathbf{Y})_{G_{\mathbf{x}}}$, and there exists a c-component $\mathbf{C}_i$ of $G^{\Delta}_{[An(\mathbf{A}_i)]}$, such that for every $P(\mathbf{V};\ \mathbf{z})$, $\mathbf{Z} \in Z^k \in Z$ at least one of the following three conditions occur:

(i) $\mathbf{A}_i \setminus {}^{k} \neq \emptyset$, that is, at least one variable in $\mathbf{A}_i$ has a different mechanism in ^{k}, or

(ii) $\mathbf{A}_i \setminus \mathbf{Z} \neq \emptyset$, meaning at least one of the variables in $\mathbf{A}_i$ have been intervened by $\mathbf{z}$ in ^{k}, or

(iii) there exists some $\mathbf{T}_i$ s.t. $\mathbf{A}_i \quad \mathbf{T}_i \quad \mathbf{C}_i$, $G_{\mathbf{z}[\mathbf{T}_i]}$ has a single c-component.

Notice that if $\mathbf{A}_i$ does not satisfy (i) and (ii) but (iii), $G_{\mathbf{z}}[\mathbf{A}_i]$ has a single c-component for every $\mathbf{Z} \in Z^k \in Z$. Since $Q[\mathbf{A};\ \mathbf{x}]$ could not be derived in $T_{P(\mathbf{V};\ \mathbf{z})}$, the expansion must

have ended in a c-factor $\mathbf{T}_i$ with a single c-component (so that the -operator cannot be used) and where every variable in an ancestor of $\mathbf{A}_i$ (so that -operator cannot be applied).

Let H be a minimal subgraph of $G^{\Delta}[\mathbf{C}_i]$ such that every edge (directed or bidirected) that can be removed without changing the fact that each $\mathbf{T}_i$ is a single c-component and the ancestral relationships between the variables, have been removed. Then we can verify that H is an s-Thicket [50, Def. 4] for $P(\mathbf{a}_i \mid do(\mathbf{v} \cap \mathbf{a}_i)) = Q[\mathbf{A}_i] = Q[\mathbf{A}_i; \mathbf{x}=\mathbf{x}]$. relative to G and Z^{0} where all non-atomic interventions in Z are replaced with atomic ones. To witness, we argue each part of the definition for s-Thicket:

- H is a minimal subgraph of G made of a single c-component,
- by conditions (i) and (ii) we have $^{k} \setminus \mathbf{R}$; and $\mathbf{Z} \setminus \mathbf{R}$; for every $\mathbf{Z}$ s.t. $_{\mathbf{Z}}=\mathbf{z}$ in Z^{0}.
- Also for every $\mathbf{Z}$, by condition (iii) there exists $\mathbf{T}_i$ such that $\langle \mathbf{T}_i; \mathbf{A}_i \rangle$ is a hedge [50].
- H's root set $\mathbf{R}$ (the variables in H without any child) is exactly $\mathbf{A}_i$, then

$$\mathbf{R} \quad An(\mathbf{A}_i)_{G_{\overline{\mathbf{V}n(\mathbf{X}[An(\mathbf{A})n\mathbf{A})}}} \quad An(\mathbf{A})_{G_{\overline{\mathbf{V}n(\mathbf{X}[An(\mathbf{A})n\mathbf{A})}}}}.$$

- Finally, for every hedgelet in those hedges, $\mathbf{T}_i \cap \mathbf{A}_i$ has to intersect $\mathbf{X} \cup An(\mathbf{A} \cap \mathbf{A})$.

This implies the existence of two sets of models $M^{(i)} = \{M^{i;k}\}_{k2}$, $i = 1; 2$, such that for every $\mathbf{Z} \subseteq Z^{k} \subseteq Z$ the corresponding models agree on $P^{k}(\mathbf{v} \mid do(\mathbf{z}))$, but disagree on $P(\mathbf{a}_i \mid do(\mathbf{v} \cap \mathbf{a}_i))$.

Using the latent factorization (eq. (2.12)) in the context of $M_{\mathbf{z}}$, for any conditional or stochastic intervention $_{\mathbf{Z}}$, the distribution $P(\mathbf{v}; _{\mathbf{Z}})$ is given by

$$P(\mathbf{v}; _{\mathbf{Z}}) = \sum_{\mathbf{u}} \prod_{\{i \mid V_i \in \mathbf{Z}\}} P(v_i \mid \mathbf{pa}_i; \mathbf{u}_i; _{\mathbf{Z}}) P(\mathbf{u} \cap \mathbf{u}; _{\mathbf{Z}}) \prod_{\{i \mid V_i \in \mathbf{V} \cap \mathbf{Z}\}} P(v_i \mid \mathbf{pa}_i; \mathbf{u}_i) P(\mathbf{u})$$
$$= P(\mathbf{v} \cap \mathbf{z} \mid do(\mathbf{z}_j)) \sum_{\mathbf{u} \cap \mathbf{u}} \prod_{\{i \mid V_i \in \mathbf{Z}\}} P(v_i \mid \mathbf{pa}_i; \mathbf{u}_i; _{\mathbf{Z}}) P(\mathbf{u} \cap \mathbf{u}; _{\mathbf{Z}}).$$

Given $M^{(1)}$ and $M^{(2)}$, we are free to specify any conditional or stochastic intervention $\mathbf{z}$, for any such $\mathbf{Z}$, by setting $P(V_i|\mathbf{Pa}_i; \mathbf{U}_i; \mathbf{z})$ for every $V_i \in \mathbf{Z}$ in the previous expression, and $P(\mathbf{v}; \mathbf{z})$ will be the same as well.

We conclude that $Q[\mathbf{A}_i; \mathbf{x}]$ is not transportable from Z and G^{Δ}. Moreover, $Q[\mathbf{A}; \mathbf{x}]$ is also not transportable from the same input; were it transportable, the -operator implies that $Q[\mathbf{A}_i; \mathbf{x}]$ can be obtained from $Q[\mathbf{A}; \mathbf{x}]$, a contradiction.

Since every variable in $\mathbf{D} \cap \mathbf{Y}$ has a children in $\mathbf{D}$, we can apply lemma 12 in a topological order over $\mathbf{D} \cap \mathbf{Y}$ and conclude that $Q[\mathbf{D}; \mathbf{x}]$ is transportable if and only if $P_{\mathrm{d}n_{\mathbf{Y}}} Q[\mathbf{D}; \mathbf{x}] = P(\mathbf{y}; \mathbf{x})$ is transportable.

Efficiency. In terms of single operations, C-INFER and its subroutines require computations such as getting the set of ancestors or c-components in a subgraph of G. These operations can be done in $O(n + m)$ time.

For the *-TR* task, each step in GENQUERYTREE is executed at most once, hence it runs in time $O(n + m)$.

The number of target factors is at most n, which is also the number of searches over any particular d-tree, one for each of the c-factors of the query. For each search in each d-tree, the and -operators are applied at most once. Moreover, every time or -operators are used, the number of variables in the factor decreases at least in one element. It follows that a search takes at most n steps. Then, the time GENINPUTTREE takes is $O(n^2(n + m)p)$.

Checking if a particular target c-factor is a leaf in a particular d-tree can be implemented in constant time. Then, MAPFACTORS takes at most $O(np)$ time.

Finally, COMPOSEQUERY has to process at most n paths (one for each target c-factor) of at most $2n$ length each. Then, it finishes in time $O(n^2)$.

As the procedures are applied one after another, the overall running time is $O(n^2(n + m)p)$, which is the maximum.

Notice that for *obs-ID*, $p = 1$ because there is only the observational distribution in P, hence the running time for that task is $O(n^2(n + m))$. □

C.2 Statistical Transportability

Theorem 8 (, , and -operators soundness and completeness for *s-TR*)**.** *Given a causal inference task with signature $I_{s\text{-}TR}$, the query Q is transportable from* P *and* G^{Δ} *if and only if* C-INFER *finds a mapping using the* , , *and* -*operators. Moreover, the task is decided in* $O(n^2(n+m))$ *time, where* $n = |\mathbf{V}|$ *and* m *is the number of edges in* G*.*

Proof. **Soundness.** The soundness follows from the validity of the , , and cftree operators.

Completeness. For *s-TR* the q-tree first uses the -operator to define the query in terms of an unconditional c-factor $Q[\mathbf{A}\,k\,\mathbf{H}]$. If $Q[\mathbf{A}\,k\,\mathbf{H}]$ is not transportable, then Q is not transportable by theorem 21 (considering an empty intervention). Then, $Q[\mathbf{A}\,k\,\mathbf{H}]$ is always branched into two subtrees. The task succeeds if any of the two sub-roots can be computed from the corresponding subtree. In particular, C-INFER fails if there exists a c-factor $Q[\mathbf{A}_i\,k\,\mathbf{H}_i]$ corresponding to a marginalized c-component of $Q[\mathbf{A}\,[\,(\mathbf{W}\setminus\mathbf{H})\,k\,\mathbf{H} n \mathbf{W}]$.

The non-transportability of $Q[\mathbf{A}_i\,k\,\mathbf{H}_i]$ implies the non-transportability of $Q[\mathbf{A}_i\,[\,(\mathbf{W}\setminus\mathbf{H})\,k\,\mathbf{H}\,n\,\mathbf{W}]$. This is because if the latter were transportable, $Q[\mathbf{A}_i\,k\,\mathbf{H}_i]$, that is always computable from $Q[\mathbf{A}\,[\,(\mathbf{W}\setminus\mathbf{H})\,k\,\mathbf{H}\,n\,\mathbf{W}]$ by -operator, would be transportable as well.

Moreover, the non-transportability of $Q[\mathbf{A}\,[\,(\mathbf{W}\setminus\mathbf{H})\,k\,\mathbf{H}\,n\,\mathbf{W}]$ implies the non-transportability of $Q[\mathbf{A}\,k\,\mathbf{H}]$ by lemma 12.

Then, we just need to prove the non-transportability of $Q[\mathbf{A}_i\,k\,\mathbf{H}_i]$ in order to prove the non-transportability of the query. This c-factor cannot be derived from $T_{P(\mathbf{W})}$ only if $\mathbf{A}_i$ is not observed in $P(\mathbf{W})$, i.e., $\mathbf{A}_i\,n\,\mathbf{W} \neq \emptyset$. On the other hand, it is not derivable from $T_{P(\mathbf{V})}$ only if $(\mathbf{A}_i\,[\,\mathbf{H}_i)\setminus \neq \emptyset$, that is, there are differences in a subset $\mathbf{A}' \subseteq \mathbf{A}_i\,[\,\mathbf{H}_i$. Notice that

$$Q[\mathbf{A}_i\,k\,\mathbf{H}_i] = \sum_{\mathbf{h}_i} Q[\mathbf{A}_i,\mathbf{h}_i] = \sum_{\mathbf{h}_i}\sum_{\mathbf{u}(\mathbf{A}_i,\mathbf{H}_i)}\prod_{V_i\in\mathbf{A}_i[\mathbf{H}_i} P(v_i \mid \mathbf{pa}_i,\mathbf{u}_i)P(\mathbf{u}(\mathbf{A}_i,\mathbf{H}_i)); \quad \text{(C.14)}$$

and the factors for each $V_i \in \mathbf{A}^0$ may differ between and . We can create two pairs of models such that $Q[\mathbf{A}_i k \mathbf{H}_i]$ match but $Q\ [\mathbf{A}_i k \mathbf{H}_i]$ does not match. Then, $Q\ [\mathbf{A}_i k \mathbf{H}_i]$ would be non-transportable, which in turns implies the non-transportability of $Q\ [\mathbf{A}_i k \mathbf{H}_i]$ (by lemma 12).

Efficiency. Atomic operations in C-INFER and its subroutines (computing ancestors, c-components, subgraphs) take $O(n + m)$ time.

For the *s-TR* task, GENQUERYTREE generates $O(n)$ nodes, so it takes at most $O(n(n+m))$ time.

Because of the number of target factors, the number of searches over any particular d-tree is also $O(n)$. For each search in each d-tree, the -operator is applied at most once. Moreover, every time or -operators are used, the number of variables in the factor decreases at least in one element. It follows that a search takes at most n steps. Then, the time GENINPUTTREE takes is $O(n^2(n + m))$.

Checking if a particular target c-factor is a leaf in a particular d-tree can be implemented in constant time. Then, MAPFACTORS takes at most $O(n)$ time.

Finally, COMPOSEQUERY has to process at most $O(n)$ paths (one for each target c-factor) of at most $2n$ length each. Then, it finishes in time $O(n^2)$.

As the procedures are applied one after another, the overall running time is $O(n^2(n + m))$, which is the maximum. □

C.3 Recoverability from Selection Bias

C.3.1 Lemmas for Model Parametrization

In order to prove non-recoverability, we need to construct structural causal models that serve as counter-examples to the recoverability of the causal effect. The following lemmas are useful to construct such models. The first one, lemma 13 licenses the the direct specification of the conditional distributions of any variable given its parents, in accordance to the causal diagram G.

Lemma 13 (Family Parametrization)**.** *Let G be a causal diagram over a set $\mathbf{V}$ of n variables. Consider also, a set of conditional distributions $P(v_i \mid \mathbf{pa}_i); 1 \leq i \leq n$ such that $\mathbf{Pa}_i$ is the set of nodes in G from which there are outgoing edges pointing into V_i. Then, there exists a model M compatible with G that induces $P(\mathbf{v}) = \prod_{i=1}^{n} P(v_i \mid \mathbf{pa}_i)$.*

Proof. (By construction) For every V_i define any ordering on the values of its domain, and let $v_i^{(j)}$ refer to the j^{th} value in that order. Also, define a continuous unobservable variable $\mathbf{U}_i \sim U[0, 1]$ (uniformly distributed in the interval $[0, 1]$) for every variable $V_i \in \mathbf{V}$. Then, construct a structural causal model $M = \langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{u}) \rangle$ where:

- $\mathbf{V}$ is the same set of observables in G
- $\mathbf{U} = \bigcup_{i=1}^{n} \mathbf{U}_i'$
- $F = \left\{ f_i(\mathbf{pa}_i, \mathbf{u}_i) = \inf_j \left[\sum_{k=1}^{j} P(v_i^{(j)} \mid \mathbf{pa}_i) > \mathbf{u}_i \right], 1 \leq i \leq n \right\}$
- $\mathbf{U}_i \sim U[0, 1], 1 \leq i \leq n$

For every variable, V_i, given a particular configuration of $\mathbf{Pa}_i$, M simulates its value using the distribution $P(v_i \mid \mathbf{pa}_i)$. By the Markov property, the joint distribution will be equal to the product of those distributions. □

The following lemma permits the construction of a structural causal model M compatible with a causal diagram G, using another model compatible with a related, but different, causal diagram G' where some arrows in a chain of variables have the reverse direction.

Lemma 14 (Chain Reversal)**.** *Consider a causal diagram G and a probability distribution $P(\mathbf{v})$ induced by any SCM M compatible with G. If G contains a chain of vertices $R_1 \rightarrow R_2, \ldots, R_\ell$ where each node represents a binary random variable, for every $1 \leq i < \ell$ the only incoming edge into R_i comes from R_{i-1}. Then, there exists another model M' where the direction of the arrows along the chain $R_1 \rightarrow R_2, \ldots, R_\ell$ is reversed compatible with the same distribution.*

Proof. (By construction) Given M and any probability distribution $P(\mathbf{v})$ induced by it, compute the joint distribution $P(r_1, \ldots, r_n, t)$. Construct a new model M' with the same set of observable variables and identical functions for all variables but for $R_1, \ldots, R_n, T$. For those, assign the functions $f_{R_i}(r_{i-1}, U_{R_i}); 1 \leq i < n$ as in lemma 13. Also, let $f_{R_n}(U_{R_n}) = U_{R_n}, P(U_{R_n}) = P(r_n)$. By lemma 13 the sub-models composed of $R_1, \ldots, R_n, T$ in M' and M produce the exact same distribution and since the set of parents and function for every other part of the model are the same, the overall distribution is identical. □

Finally, the following lemma allows to simplify the parametrization of an arbitrarily long chain of binary variables.

Lemma 15 (Collapsible Path Parametrization)**.** *Consider a causal diagram G and a probability distribution $P(\mathbf{v})$ induced by any SCM compatible with G. If G contains a chain $W_0 \to W_1 \to \cdots \to W_k$, where each W_i represents a binary random variable, for every $1 \leq i \leq k$ the only incoming edge to W_i is from W_{i-1}, and every conditional distribution $P(w_i \mid w_{i-1}) = p$, $P(w_i \mid \overline{w_{i-1}}) = q$, for some $0 < p, q < 1$. Then, the conditional distribution $P(w_k \mid w_0) = \frac{q - (p-1)(p-q)^k}{q-p+1}$, $P(w_k \mid \overline{w_0}) = \frac{q - q(p-q)^k}{q-p+1}$.*

Proof. Since $W_0, \ldots, W_k$ is a chain, the value of W_k is a function of W_0 when all other $W_1, \ldots, W_{k-1}$ are marginalized. All W_i, $1 \leq i \leq k$ are independent of any other variable given W_0. Therefore, the distribution $P(w_k \mid w_0)$ is equal to $\prod_{i=1}^{k-1} \sum_{k} P(w_i \mid w_{i-1})$, because any other variable can be removed from any product in this expression and summed out. This distribution can be calculated as the product of 2x2 matrices corresponding to the conditional distributions $P(w_i \mid w_{i-1})$ when encoded as

$$W_M = \begin{bmatrix} p & q \\ 1-p & 1-q \end{bmatrix}. \tag{C.15}$$

The product of k of such matrices is readily available if W_M is decomposed using its

eigenvalues $\{1, p - q\}$ and eigenvectors $\left\{\left[\frac{q}{(1-p)}, 1\right], [1, 1]\right\}$:

$$P(w_k \mid w_0) = \prod_{i=1}^{k} P(w_i \mid w_{i-1}) = (W_M)^k$$
$$= \begin{bmatrix} \frac{q - (p-1)(p-q)^k}{q-p+1} & \frac{q - q(p-q)^k}{q-p+1} \\ 1 - \frac{q - (p-1)(p-q)^k}{q-p+1} & 1 - \frac{q - q(p-q)^k}{q-p+1} \end{bmatrix} \tag{C.16}$$

□

The following lemma extends a result from [102], to the context of recoverability.

Lemma 16. *Let $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be two disjoint sets of variables and $\mathbf{T} \subseteq \mathbf{V}$ another set of variables. If $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in G from $P(\mathbf{v} \mid S=1), P(\mathbf{t})$, then $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in the graph resulted from adding a directed or bidirected edge to G. Equivalently, if $P(\mathbf{y} \mid do(\mathbf{x}))$ is recoverable in G from $P(\mathbf{v} \mid S=1), P(\mathbf{t})$, then $P(\mathbf{y} \mid do(\mathbf{x}))$ is still recoverable in the graph resulted from removing a directed or bidirected edge from G.*

Proof. The proof is analogous to the proof of lemma 8 in [102] with the extra consideration of selection bias. For any $V_i \in \mathbf{V} \cup \{S\}$ let $\mathbf{Pa}_i^+$ be defined as the set of observable and unobservable parents of V_i in G.

If $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in G, then there exist two models with the same causal graph G, M_1 and M_2 such that

$$P^{M_1}(\mathbf{v} \mid S=1) = P^{M_2}(\mathbf{v} \mid S=1) > 0,$$
$$P^{M_1}(\mathbf{t}) = P^{M_2}(\mathbf{t}) > 0,$$
$$\text{and } P_{\mathbf{x}}^{M_1}(\mathbf{y}) \neq P_{\mathbf{x}}^{M_2}(\mathbf{y}) \tag{C.17}$$

where

$$P^{M_k}(\mathbf{v}|S=1) = \sum_{\mathbf{U}} \frac{P(S=1|\mathbf{pa}_S^+)}{P(S=1)} \prod_{V_i \in \mathbf{V}} P(v_i|\mathbf{pa}_i^+)P(\mathbf{u}) \qquad ; k = 1,2 \tag{C.18}$$

$$P^{M_k}(\mathbf{t}) = \sum_{\mathbf{U}} \sum_{\mathbf{V} \cap \mathbf{T}} \prod_{V_i \in \mathbf{V}} P(v_i \mid \mathbf{pa}_i^+)P(\mathbf{u}) \qquad ; k = 1,2 \tag{C.19}$$

For a graph G' with extra edges added to G, we can always construct new models in such a way that the added edges are ineffective.

(i) Let G' be the graph identical to G except with an extra edge $W \to V_j$. Then $P(\mathbf{v} \mid S=1)$ and $P(\mathbf{t})$ decompose as

$$P(\mathbf{v}|S=1) = \sum_{\mathbf{U}} \frac{P(S=1|\mathbf{pa}_S^+)}{P(S=1)} P(v_j|\mathbf{pa}_{V_j}^+, w) \prod_{\substack{V_i \in \mathbf{V}; \\ V_i \neq V_j}} P(v_i|\mathbf{pa}_i^+)P(\mathbf{u}) \tag{C.20}$$

$$P(\mathbf{t}) = \sum_{\mathbf{U}} \sum_{\mathbf{V} \cap \mathbf{T}} P(v_j|\mathbf{pa}_{V_j}^+, w) \prod_{\substack{V_i \in \mathbf{V}; \\ V_i \neq V_j}} P(v_i|\mathbf{pa}_i^+)P(\mathbf{u}) \tag{C.21}$$

We construct two models M_1' and M_2' with causal graph G' as:

$$P^{M_k'}(v_i|\mathbf{pa}_i^+) = P^{M_k}(v_i|\mathbf{pa}_i^+) \qquad ; i \neq j; k=1,2 \tag{C.22}$$

$$P^{M_k'}(S=1|\mathbf{pa}_S^+) = P^{M_k}(S=1|\mathbf{pa}_S^+) \qquad ; k=1,2 \tag{C.23}$$

$$P^{M_k'}(v_j|\mathbf{pa}_{V_j}^+, w) = P^{M_k}(v_j|\mathbf{pa}_{V_j}^+) \qquad ; k=1,2 \tag{C.24}$$

$$P^{M_k'}(\mathbf{u}) = P^{M_k}(\mathbf{u}) \qquad ; k=1,2 \tag{C.25}$$

Clearly if the pair (M_1, M_2) satisfies (C.17) so does (M_1', M_2'). Hence $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in G'.

(ii) Let G' be the graph identical to G except with an extra edge $V_l \leftrightarrow V_j$. Then $P(\mathbf{v} \mid S=1)$

and $P(\mathbf{t})$ decompose as

$$P(\mathbf{v}|S=1) = \sum_{u'} P(u') \sum_{\mathbf{U}} \frac{P(S=1|\mathbf{pa}_S^+)}{P(S=1)} P(v_j|\mathbf{pa}_{V_j}^+, u') P(v_l|\mathbf{pa}_{V_l}^+, u') \prod_{V_i \in \mathbf{V}; V_i \notin V_j; V_i \notin V_l} P(v_i|\mathbf{pa}_i^+) P(\mathbf{u}) \tag{C.26}$$

$$P(\mathbf{t}) = \sum_{u'} P(u') \sum_{\mathbf{U}} \sum_{\mathbf{V} \cap \mathbf{T}} P(v_j|\mathbf{pa}_{V_j}^+, u') P(v_l|\mathbf{pa}_{V_l}^+, u') \prod_{V_i \in \mathbf{V}; V_i \notin V_j; V_i \notin V_l} P(v_i|\mathbf{pa}_i^+) P(\mathbf{u}) \tag{C.27}$$

Where U' represents a new unobservable variable. We construct two models M_1' and M_2' with causal graph G' as:

$$P^{M_k'}(v_i|\mathbf{pa}_i^+) = P^{M_k}(v_i|\mathbf{pa}_i^+) \quad ; i \notin j; i \notin l; k = 1, 2 \tag{C.28}$$

$$P^{M_k'}(S=1|\mathbf{pa}_S^+) = P^{M_k}(S=1|\mathbf{pa}_S^+) \quad ; k = 1, 2 \tag{C.29}$$

$$P^{M_k'}(v_i|\mathbf{pa}_i^+, u') = P^{M_k}(v_i|\mathbf{pa}_i^+) \quad ; i = j, l; k = 1, 2 \tag{C.30}$$

$$P^{M_k'}(\mathbf{u}) = P^{M_k}(\mathbf{u}) \quad ; k = 1, 2 \tag{C.31}$$

Again, if the pair (M_1, M_2) satisfies (C.17), so does (M_1', M_2'). Hence $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in G'.

(iii) Let G' be the graph identical to G except with an extra edge $W \to S$. Then $P(\mathbf{t})$ is exactly the same and $P(\mathbf{v} \mid S=1)$ decomposes as

$$P(\mathbf{v}|S=1) = \sum_{\mathbf{U}} \frac{P(S=1|\mathbf{pa}_S^+, w)}{P(S=1)} \prod_{V_i \in \mathbf{V}} P(v_i|\mathbf{pa}_i^+) P(\mathbf{u}) \tag{C.32}$$

We construct two models $M_1^{\prime}$ and $M_2^{\prime}$ with causal graph $G^{\prime}$ as:

$$P^{M_k^{\prime}}(v_i|\mathbf{pa}_i^{+}) = P^{M_k}(v_i|\mathbf{pa}_i^{+}) \qquad ; k = 1,2 \tag{C.33}$$

$$P^{M_k^{\prime}}(S=1|\mathbf{pa}_S^{+}, w) = P^{M_k}(S=1|\mathbf{pa}_S^{+}) \qquad ; k = 1,2 \tag{C.34}$$

$$P^{M_k^{\prime}}(\mathbf{u}) = P^{M_k}(\mathbf{u}) \qquad ; k = 1,2 \tag{C.35}$$

Since $P(S=1) = \sum_{\mathbf{pa}_S^{+}} P(S=1 \mid \mathbf{pa}_S^{+}) P(\mathbf{pa}_S^{+})$, that distribution will remain the same. Then, if pair (M_1, M_2) satisfies (C.17) so does $(M_1^{\prime}, M_2^{\prime})$. Hence $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in $G^{\prime}$.

(iv) Let $G^{\prime}$ be the graph identical to G except with an extra edge $V_j \leftrightarrow S$. Then $P(\mathbf{v} \mid S=1)$ and $P(\mathbf{t})$ decompose as

$$P(\mathbf{v}|S=1) = \sum_{u^{\prime}} P(u^{\prime}) \sum_{\mathbf{U}} \frac{P(S=1|\mathbf{pa}_S^{+}, u^{\prime})}{P(S=1)} P(v_j|\mathbf{pa}_{V_j}^{+}, u^{\prime}) \prod_{V_i \in \mathbf{V}, V_i \neq V_j} P(v_i \mid \mathbf{pa}_i^{+}) P(\mathbf{u}) \tag{C.36}$$

$$P(\mathbf{t}) = \sum_{u^{\prime}} P(u^{\prime}) \sum_{\mathbf{U}} \sum_{\mathbf{V} \cap \mathbf{T}} P(v_j|\mathbf{pa}_{V_j}^{+}, u^{\prime}) \tag{C.37}$$

$$\prod_{V_i \in \mathbf{V}, V_i \neq V_j} P(v_i|\mathbf{pa}_i^{+}) P(\mathbf{u}) \tag{C.38}$$

Where $U^{\prime}$ represents a new unobservable variable. We construct two models $M_1^{\prime}$ and $M_2^{\prime}$ with causal graph $G^{\prime}$ as:

$$P^{M_k^{\prime}}(v_i|\mathbf{pa}_i^{+}) = P^{M_k}(v_i|\mathbf{pa}_i^{+}) \qquad ; i \neq j, k = 1,2 \tag{C.39}$$

$$P^{M_k^{\prime}}(S=1|\mathbf{pa}_S^{+}, u^{\prime}) = P^{M_k}(S=1|\mathbf{pa}_S^{+}) \qquad ; k = 1,2 \tag{C.40}$$

$$P^{M_k^{\prime}}(v_j|\mathbf{pa}_{V_j}^{+}, u^{\prime}) = P^{M_k}(v_j|\mathbf{pa}_{V_j}^{+}) \qquad ; k = 1,2 \tag{C.41}$$

$$P^{M_k^{\prime}}(\mathbf{u}) = P^{M_k}(\mathbf{u}) \qquad ; k = 1,2 \tag{C.42}$$

Again, if the pair $(M_1; M_2)$ satisfies (C.17), so does $(M_1^{\prime}; M_2^{\prime})$. Hence $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in $G^{\prime}$.

□

Theorem 19 (Necessary Graphical Condition for *sb-ID*)**.** *A sb-ID task with signature* $I = \langle P(\mathbf{y} \mid do(\mathbf{x}); S = 1); \{P(\mathbf{V} \mid S = 1)\}; G \rangle$ *is solvable only if* $(\mathbf{Y} \not\perp\!\!\!\perp S)_{G^{pbd}_{\mathbf{XY}}}$.

Proof. Suppose the stated condition is satisfied, there exists an active path $\bar{p}$ between some $Y^{\prime} \in \mathbf{Y}$ and S in $G^{pbd}_{\mathbf{XY}}$.

Without loss of generality, let $Y^{\prime}$ be the element in $\mathbf{Y}$ connected to S with the shortest active path $\bar{p}$. Let $\mathbf{X}^{\prime}$ be the variables in $\mathbf{X}$ that lie in $\bar{p}$. Let $G^{\prime}$ be the subgraph of G that contains the same set of variables but only the edges in $\bar{p}$. We will show that $P_{\mathbf{x}^{\prime}}(y^{\prime})$ is not recoverable from $P(\mathbf{v} \mid S{=}1)$ in $G^{\prime}$, when this is the case it's easy to check that $P_{\mathbf{x}}(\mathbf{y}^{\prime})P_{\mathbf{x}^{\prime}}(\mathbf{y}^{\prime})$ is also not recoverable in $G^{\prime}$. Then by lemma 16 this will imply in turn that $P(\mathbf{y} \mid do(\mathbf{x}))$ is not recoverable in G.

Let $\mathbf{V}$ represent all variables in the graph except for the selection mechanism S, and let $Q = P(y^{\prime} \mid do(\mathbf{x}^{\prime}))$. We construct two SCMs M_1 and M_2 compatible with $G^{\prime}$, that induce probability distributions $P^{M_1}(\mathbf{v} \mid S{=}1)$ and $P^{M_2}(\mathbf{v} \mid S{=}1)$, respectively, such that

$$P^{M_1}(\mathbf{v} \mid S{=}1) = P^{M_2}(\mathbf{v} \mid S{=}1) \tag{C.43}$$

$$Q^{M_1} \neq Q^{M_2} \tag{C.44}$$

Let M_1 be compatible with $G^{\prime}$ and M_2 with $G^{\prime}_{\overline{S}}$, enforcing $(\mathbf{V} \not\perp\!\!\!\perp S)_{P^{M_2}}$. Without loss of generality, all variables are assumed to be binary. The construction parametrizes P^{M_1} through its factors (as in lemma 13) and then parametrizes P^{M_2} to enforce (C.43). As a consequence, $P^{M_2}(\mathbf{v}) = P^{M_2}(\mathbf{v} \mid S{=}1) = P^{M_1}(\mathbf{v} \mid S{=}1)$.

In the sequel we consider every possible form in which $\bar{p}$ could manifest in G:

case 1 $Y^{\prime} \in Pa_S$

The causal effect in M_2:

$$\begin{aligned} Q^{M_2} &= P^{M_2}(y') = P^{M_1}(y' \mid S=1) \\ &= \frac{P^{M_1}(y', S=1)}{\sum_{y''} P^{M_1}(y'', S=1)} \\ &= \frac{P^{M_1}(S=1 \mid y') P^{M_1}(y')}{P^{M_1}(S=1 \mid y') P^{M_1}(y') + P^{M_1}(S=1 \mid \overline{y'}) P^{M_1}(\overline{y'})} \end{aligned}$$

Using lemma 13, let $P^{M_1}(S=1 \mid y') =$ and $P^{M_1}(S=1 \mid \overline{y'}) =$ with $0 <$; < 1 and and $P^{M_1}(y') = 1/2$. We obtain:

$$Q^{M_2} = \frac{}{+}$$

By the same reasoning, Q^{M_1} is equal to 1/2 and it is never equal to Q^{M_2} given this parametrization.

case 2 There is a directed path $\overline{p}$ from Y' to S.

Let R be the parent of S in such path and let $\mathbf{W}$ be the set of variables in the path from Y' to R. Note that even if $\mathbf{W} \setminus \mathbf{X} \neq ;$, $Q^{M_2} = P(y')$, then, similar to the previous case:

$$Q^{M_2} = P^{M_1}(y' \mid S=1) = \frac{P^{M_1}(y', S=1)}{P^{M_1}(S=1)}$$

The numerator can be rewritten as:

$$\begin{aligned} P^{M_1}(y', S=1) &= \sum_R P^{M_1}(y', r, S=1) \\ &= \sum_R P^{M_1}(y') P^{M_1}(r \mid y') P^{M_1}(S=1 \mid r) \end{aligned}$$

Factorizing the denominator analogously, Q^{M_2} becomes:

$$Q^{M_2} = \frac{P^{M_1}(y^0) \sum_R P^{M_1}(r \mid y^0) P^{M_1}(S=1 \mid r)}{\sum_{Y^0} P^{M_1}(y^0) \sum_R P^{M_1}(r \mid y^0) P^{M_1}(S=1 \mid r)}$$

Use lemma 15 to set $P^{M_1}(r \mid y^0) = 1/2 + \varepsilon = 2/5; P^{M_1}(r \mid \overline{y^0}) = 1/2 - \varepsilon = 2/5$ where $\varepsilon = (1/5)^k$ (using $p = 3/5; q = 2/5$). Also let $P^{M_1}(S=1 \mid r) = 2/3$ and $P^{M_1}(S=1 \mid \bar{r}) = 1/2$ and $P^{M_1}(y^0) = 1/2$. This parametrization leads to $Q^{M_2} = 1/2 + \varepsilon = 14$ and $Q^{M_1} = 1/2$ which are never equal.

case 3 The path $\bar{p}$ connecting Y^0 and S goes through an ancestor of both.

Let N be the common ancestor of Y^0 and S in $\bar{p}$. Let R be the parent of S and Q the parent of Y^0 in the mentioned path. Let $\mathbf{W_1}$ and $\mathbf{W_2}$ be the nodes in the paths from N to Q and from N to R respectively. Consider an equivalent graph G'' where the arrows in the subpath from N to Q are reversed. Any model constructed for G'' can be translated to a model compatible with G^0 using 14. Again we have $Q^{M_2} = P^{M_2}(y^0)$ Following the same derivation as in *case 2* yields:

$$Q^{M_2} = \frac{P^{M_1}(y^0; S=1)}{\sum_{Y^0} P^{M_1}(y^0; S=1)}$$

The numerator of the last expression can be rewritten as:

$$\begin{aligned} P^{M_1}(y^0; S=1) &= \sum_Q P^{M_1}(y^0; q; S=1) \\ &= \sum_Q P^{M_1}(y^0 \mid q) P^{M_1}(q) P^{M_1}(S=1 \mid q) \end{aligned}$$

By rewriting the denominator similarly, and following an analogous process for Q^{M_1},

we have:

$$Q^{M_1} = \sum_{} P^{M_1}(y' \mid q) P^{M_1}(q)$$

$$Q^{M_2} = \frac{\sum_{Q} P^{M_1}(y' \mid q) P^{M_1}(q) P^{M_1}(S=1 \mid q)}{\sum_{Y';Q} P^{M_1}(y' \mid q) P^{M_1}(q) P^{M_1}(S=1 \mid q)}$$

Lemma 15 can be employed to set $P^{M_1}(r \mid q) = 1=2 + =2; P^{M_1}(r \mid \bar{q}) = 1=2 =2$ where $= (1=5)^k$ (using $p = 3=5; q = 2=5$). Define $P^{M_1}(S=1 \mid r) = 2=3$ and $P^{M_1}(S=1 \mid \bar{r}) = 1=2$. Calculate $P^{M_1}(S=1 \mid q)$ as $\sum_{R} P^{M_1}(r \mid q) P^{M_1}(S=1 \mid r)$. Also let $P^{M_1}(y' \mid q) = 3=4; P^{M_1}(y' \mid \bar{q}) = 1=2$, finally $P^{M_1}(q) = 1=2$. This parametrization leads to:

$$Q^{M_1} = \frac{5}{8} \qquad\qquad Q^{M_2} = \frac{5}{8} + \frac{}{56}$$

which are never equal.

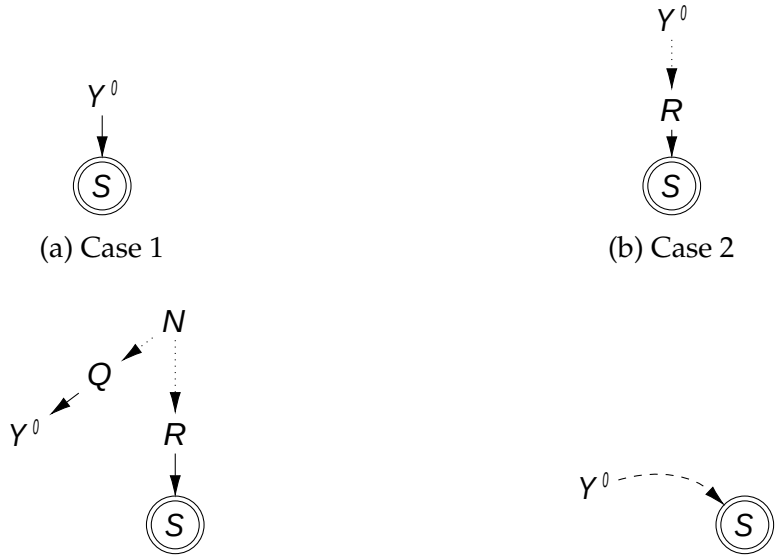
case 4 $\bar{p}$ is a confounding path between Y' and S consisting of unobservable variables.

The models for this case can be constructed as in *case 3,* then moving the variables in the in the path $Q \quad R$ (included) from the set of observables to the set of unobservables.

□

Theorem 18 (-, -, -, $'$-, and -operators soundness and completeness for sb-ID)**.** *Given a causal inference task with signature* $I_{sb\text{-}ID} = \langle Q = P(\mathrm{y} \mid do(\mathrm{x})); \{P(\mathbf{V} \mid S = 1)\}; \{G\} \rangle$, *the query Q is recoverable from* $P(\mathbf{V} \mid S = 1)$ *and G if and only if* C-INFER *finds a mapping using the , , and* $'$*-operators. Moreover, the task is decided in* $O(n^2(n + m))$ *time, where* $n = |\mathbf{V}|$ *and m is the number of edges in G.*

Proof. **Soundness.** The soundness follows from the validity of the , , , and $'$ cftree operators.



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

Figure C.1: Graphical representation of the cases stated in theorem 19

Completeness. *C-Infers* fails at a *sb-ID* task if there exists some c-factor $Q[\mathbf{D}_i \mid do(\mathbf{x})]$ in the q-tree that could not be mapped from any of the available d-trees.

This means that the expansion of $T_{P(\mathbf{V}_{jS=1})}$ towards $Q[\mathbf{D}_i \mid do(\mathbf{x})]$ must have stopped at some minimal c-factor $Q[\mathbf{C}]$ or $Q[\mathbf{C} \mid S]$, where $\mathbf{D}_i \not\subseteq \mathbf{C}$. If it stopped at $Q[\mathbf{C}]$, notice that $Q[\mathbf{C}]$ must be computable also from $Q[\mathbf{V}]$ hence the same query would fail even if the input was $P(\mathbf{V})$. By the completeness of C-INFER for *obs-ID*, we conclude that Q is also not identifiable in this case.

If the expansion stopped at some c-factor $Q[\mathbf{C} \mid S]$ in $T_{P(\mathbf{V}_{jS=1})}$, it must be the case that the $,$ and $^{(}$ operators cannot be used to obtain a smaller c-factor that contains $\mathbf{D}_i$. Because the $-$-operator does not apply, it follows that $\mathbf{C} \subseteq An(\mathbf{D}_i \cup \{S\})$. Similarly, for the $\cap$ and $^{(}$ operators not to apply it must be the case that every c-component of $G[\mathbf{C} \cup \{S\}]$ contains an ancestor of S. We will show that the existence of $Q[\mathbf{C} \mid S = 1]$ implies the existence of an active path $\overline{p}^{(}$ from S to some $D^{(} \in \mathbf{D}_i$ in $G[\mathbf{C} \cup \{S\}]$. Let $\mathbf{C}_i$ be the c-component of $G[\mathbf{C} \cup \{S\}]$ such that $\mathbf{D}_i \subseteq \mathbf{C}_i$:

- If $\mathbf{C}_i \setminus An(\mathbf{D}_i) \setminus An(S) \neq \emptyset$, let $C^{(}$ be any element in that intersection. Then $\overline{p}^{(}$ is the directed path from $C^{(}$ to S and from $C^{(}$ to some $D^{(} \in \mathbf{D}_i$ (possibly of length 0 if

$C' \in \mathbf{D}_i$). If the segment $C' \to \cdots \to S$ contains any variable in $X' \in \mathbf{X}$, then the outgoing edge of X' exists in $G^{pbd}_{\mathbf{XY}}$ unless there is some $Y' \in \mathbf{Y}$ that is a descendant of X' through that edge.

- Else $(\mathbf{C}_i \setminus An(S)) \cap An(\mathbf{D}_i)$ has to be non-empty, let C' be any element in that set connected with some $W \in \mathbf{C}_i \setminus An(\mathbf{D}_i)$ with a bidirected arrow. Such C' and W must exist for $\mathbf{C}_i$ to be a c-component in $G[\mathbf{C} \cup \{S\}]$. Then $\overline{p'}$ is formed by the subpaths from $C' \to \cdots \to S$ (possibly of length zero if $C' = S$), $C' \leftrightarrow W$ and $W \to \cdots \to D'$, for some $D' \in \mathbf{D}_i$.

Since $Q[\mathbf{D}_i \mid do(\mathbf{x})]$ is a target c-factor in T_Q, all its elements, including D', has a directed path from to some $Y' \in \mathbf{Y}$ (possibly of length zero if $D' \in \mathbf{Y}$), which is not intersected by any $\mathbf{X}$. As a consequence, we can construct an active path $\overline{p}$ in $G^{pbd}_{\mathbf{XY}}$ formed by $\overline{p'}$ and the directed path from D' to Y'. This implies $(\mathbf{Y} \not\perp\!\!\!\perp S)_{G^{pbd}_{\mathbf{XY}}}$ and by theorem 19 $Q = P(\mathrm{y} \mid do(\mathrm{x}))$ is not recoverable from $\mathsf{P} = \{P(\mathbf{V} \mid S{=}1)\}$.

Efficiency. In terms of single operations, C-INFER and its subroutines require computations such as getting the set of ancestors or c-components in a subgraph of G. These operations can be done in $O(n + m)$ time.

For the *sb-ID* task, each step in GENQUERYTREE is executed at most once, hence it runs in time $O(n + m)$.

The number of target factors is at most n, which is also the number of searches over any particular d-tree, one for each of the c-factors of the query. As there is only d-tree, the -operator is applied at most once per target factor. Moreover, every time , , and $'$-operators are used, the number of variables in the factor decreases at least in one element. It follows that a search takes at most n steps. Then, the time GENINPUTTREE takes is $O(n^2(n + m))$.

Checking if a particular target c-factor is a leaf in a particular d-tree can be implemented in constant time. Then, MAPFACTORS takes at most $O(np)$ time.

Finally, COMPOSEQUERY has to process at most n paths (one for each target c-factor) of at most $2n$ length each. Then, it finishes in time $O(n^2)$.

As the procedures are applied one after another, the overall running time is $O(n^2(n + m))$, which is the maximum. □

Finally, we refer to the soundness result for *sbt-ID*.

Theorem 20 (-, -, -, ^{0}-, -, ^{0}-, and -operators soundness *sbt-ID*)**.** *Given a causal inference task with signature* $I_{sbt\text{-}ID} = \langle P(\mathbf{y} \mid do(\mathbf{x})); \{P(\mathbf{V} \mid S = 1); P(\mathbf{T})\}; \{G\}\rangle$*, the query is recoverable from* $P(\mathbf{V} \mid S = 1)$*,* $P(\mathbf{T})$ *and* G *if* C-INFER *finds a mapping using the* , , , ^{0} *and* -*operators. Moreover, the process takes* $O(n^2(n + m))$ *time, where* $n = |\mathbf{V}|$ *and* m *is the number of edges in* G*.*

Proof. **Soundness.** The soundness follows from the validity of the , , , , ^{0}, and ^{0} cftree operators.

Efficiency. The analysis is the same as for the proof of theorem 18, except for the fact that there are two d-trees in the input, hence *GenInputTree* takes time $O(2n^2(n + m)) = O(n^2(n + m))$ and the overall running time is the same. □

C.4 Conditional Queries

First, we introduce some auxiliary lemmata.

Lemma 17. *Let* $A; B$ *and* C *be binary random variables causally related as given by the chain* $A \to B \to C$*. And suppose* $P(B = 1 \mid A = 1) =$ *and* $P(B = 1 \mid A = 0) = 1$ *, for some* $2 [0; 1]$*. Then, for any such that* $|1=2$ $|1=2$ *there is always a function* f_C *such that* $P(C = 1 \mid A = 1) =$ *,* $P(C = 1 \mid A = 0) = 1$ *and* $P(A; B; C)$ *is a positive distribution.*

Proof. Let $P(C = 1 \mid B = 1) = x$ and $P(C = 1 \mid B = 0) = 1 \quad x$, then

$$= P(C = 1 \mid A = 1) = \sum_{b} P(C = 1 \mid b)P(b \mid A = 1) = 1 \qquad + x(2 \qquad 1) \qquad \text{(C.45)}$$

$$x = \frac{+ \qquad 1}{2 \qquad 1}: \qquad \text{(C.46)}$$

Since x must belong to the interval $(0; 1)$, we can bound $\quad 2\ (1 \quad ; \quad)$ if $\quad 1=2$ and $\quad 2\ (\ ; 1 \quad)$ if $\quad > 1=2$. Both conditions are satisfied when $j1=2 \quad j \quad j1=2 \quad j$ as assumed, so any solution x is a valid probability.

Then, we can define the function for C as

$$f_C = B \quad U_c; \qquad \text{(C.47)}$$

where, U_c is a binary unobservable, $P(U_c = 0) = x$ and $\quad$ is the logical xor operator. □

Lemma 18. *Let A and B be two variables in a causal graph where $A \quad A_n \quad A_{n\ 1} \quad A_1 \quad C \ ! \ B_1 \ ! \quad ! \ B_{m\ 1} \ ! \ B_m \ ! \ B$. The variables $A_1; : : : ; A_n; B_1; : : : ; B_m$ are observable, C could be observable or unobservable and $m; n$ are non-negative integers. Then we can define functions for all variables involved such that they are binary and*

$$P(a; b) = \begin{cases} \frac{1}{2} & \text{if } a = b \\ \frac{1}{2}(1 \quad) & \text{otherwise} \end{cases}; \qquad \text{(C.48)}$$

for any $\quad 2\ (0; 1)$.

Proof. First, if C is unobservable set $P(C = 1) = 1=2$, else define an unobservable U_c with $P(U_c = 1) = 1=2$ and let $f_C = U_c$. Let $\quad ; \quad 2\ (0; 1)$ be parameters to decide later.

If $n = 0$ define f_A such that $P(A = 1 \mid C = 1) = \quad , P(A = 1 \mid C = 0) = 1 \quad$. Similarly, if $m = 0$ define f_B such that $P(B = 1 \mid C = 1) = \quad , P(B = 1 \mid C = 0) = 1 \quad$.

Suppose $n > 0$, then we will define the functions for $A_1; A_2; : : : ; A_n; A$ such that $P(A_i =$

$1\ j\ C\ =\ 1)$ gets closer to as i increases. If $<\ 1=2$, set f_{A_1} such that $P(A_1\ =\ 1\ j\ C\ =\ 1)\ =$ $=(n\ +\ 1), P(A_1\ =\ 1\ j\ C\ =\ 0)\ =\ 1$ $=(n\ +\ 1)$. Then use lemma 17 to define $f_{A_i}, i\ =\ 1;:::;n$ and f_A such that $P(A_i\ =\ 1\ j\ C\ =\ 1)\ =\ i$ $=(n\ +\ 1), P(A_i\ =\ 1\ j\ C\ =\ 0)\ =\ 1$ i $=(n\ +\ 1)$ and finally $P(A\ =\ 1\ j\ C\ =\ 1)\ =\ (n\ +\ 1)$ $=(n\ +\ 1)\ =$ $, P(A\ =\ 1\ j\ C\ =\ 1)\ =\ 1$.

If $>\ 1=2$ use the same strategy but starting from $P(A_1\ =\ 1\ j\ C\ =\ 1)\ =\ 1$ $(1$ $)=(n+1)$ and decreasing as $P(A_i\ =\ 1\ j\ C\ =\ 1)\ =\ 1$ $i(1$ $)=(n\ +\ 1)$, to obtain $P(A\ =\ 1\ j\ C\ =\ 1)\ =$ 1 $(n\ +\ 1)(1$ $)=(n\ +\ 1)\ =$.

The same procedure is applied for $B_1;:::;B_m;B$ to obtain $P(B\ =\ 1\ j\ C\ =\ 1)\ =$, $P(B\ =\ 1\ j\ C\ =\ 0)\ =\ 1$.

Finally,

$$P(A\ =\ 1;B\ =\ 1)\ =\ \sum_c P(A\ =\ 1\ j\ c)P(B\ =\ 1\ j\ c)P(c) \tag{C.49}$$

$$=\ \tfrac{1}{2}[\quad +\ (1\quad)(1\quad)] \tag{C.50}$$

$$P(A\ =\ 0;B\ =\ 0)\ =\ \tfrac{1}{2}[(1\quad)(1\quad)\ +\quad] \tag{C.51}$$

$$P(A\ =\ 0;B\ =\ 1)\ =\ \tfrac{1}{2}[(1\quad)\ +\ (1\quad)] \tag{C.52}$$

$$P(A\ =\ 1;B\ =\ 0)\ =\ \tfrac{1}{2}[\ (1\quad)\ +\ (1\quad)\ ]: \tag{C.53}$$

If $<\ 1=2$ make $=\ 1$ and $=\ \frac{1\ \sqrt{1\ \ 2}}{2}$. If $1=2$, let $=$ and $=\ \frac{1\ \sqrt{2\ \ 1}}{2}$. It is just a matter of algebra to verify that $P(A;B)$ results in the intended distribution. □

Theorem 21 (C-INFER soundness and completeness w.r.t. solving conditional *obs-ID*, *g-ID*, *g-TR*, and *-TR*.)**.** *Consider a causal inference task with one of the following signatures:*

$$I_{\text{obs-IDC}}\ =\ \langle P(\mathbf{y}\ j\ do(\mathbf{x});\mathbf{z});\ \{P(\mathbf{V})\};\ \{G\}\rangle; \tag{7.1}$$

$$I_{\text{g-IDC}}\ =\ \langle P(\mathbf{y}\ j\ do(\mathbf{x});\mathbf{z});\mathsf{P};\ \{G\}\rangle; \tag{7.2}$$

$$I_{\text{g-TRC}}\ =\ \langle P\ (\mathbf{y}\ j\ do(\mathbf{x});\mathbf{z});\mathsf{P};G^{\Delta}\ \rangle;\ \textit{or} \tag{7.3}$$

$$I\ _{\text{-TRC}}\ =\ \langle P\ (\mathbf{y}\ j\ \mathbf{z};\ \mathbf{x});\mathsf{P};G^{\Delta}\ \rangle; \tag{7.4}$$

where P *matches the unconditional version of the tasks. Then, the query Q is identifiable / transportable from* P *and* G^{Δ} *if and only if* C-INFER *finds a mapping using the* , , , , *and operators. Moreover, the task is decided in* $O(n^2(n+m)p)$ *time, where* $n = |\mathbf{V}|$, *m is the number of edges in G and* $p = |\mathbf{P}|$.

Proof. As -*TR* subsumes *obs-ID, g-ID,* and *g-TR,* it suffices to prove the result for the former task.

Suppose the query has the form $P(\mathbf{y} \mid \mathbf{w}; \mathbf{x})$ and let $\mathbf{W}_{\mathbf{y}}$ $\mathbf{W}$ be the variables in the same marginalized c-component as any $\mathbf{Y}$ summing $An(\mathbf{Y})_{G_{\underline{\mathbf{X}}} \cap \mathbf{W}}$. Then, let $\mathbf{W}_{\overline{\mathbf{y}}} = \mathbf{W} \setminus \mathbf{W}_{\mathbf{y}}$.

First we argue that $(\mathbf{Y} \perp\!\!\!\perp \mathbf{W}_{\overline{\mathbf{y}}} \mid \mathbf{W}_{\mathbf{y}})$ in $G_{\underline{\mathbf{X}}\underline{\mathbf{W}_{\overline{\mathbf{y}}}}}$ and $G_{\underline{\mathbf{X}}\overline{\mathbf{W}_{\overline{\mathbf{y}}}}\underline{\mathbf{W}_{\overline{\mathbf{y}}}}}$. The separation in the latter graph is obvious since we are cutting both the incoming and outgoing edges to $\mathbf{W}_{\overline{\mathbf{y}}}$. For the first graph suppose the separation does not hold, then let π be the path between $Y \in \mathbf{Y}$ and $W \in \mathbf{W}_{\overline{\mathbf{y}}}$ that is active in $G_{\underline{\mathbf{X}}\underline{\mathbf{W}_{\overline{\mathbf{y}}}}}$ given $\mathbf{W}_{\mathbf{y}}$. Path π must not have any edge going out of $\mathbf{W}_{\mathbf{y}}$ else it would be blocked by conditioning on that set. Then, π exists in $G_{\underline{\mathbf{X}}\underline{\mathbf{W}}}$, but this would make W, by definition, part of $\mathbf{W}_{\mathbf{y}}$, a contradiction. Since the stated separation holds we can use rule 2 of -calculus to infer

$$P(\mathbf{y} \mid \mathbf{w}; \mathbf{x}) = P(\mathbf{y} \mid \mathbf{w}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}})). \tag{C.54}$$

By definition of conditional probability:

$$P(\mathbf{y} \mid \mathbf{w}; \mathbf{x}) = P(\mathbf{y} \mid \mathbf{w}_{\mathbf{y}}; \mathbf{w}_{\overline{\mathbf{y}}}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}})) \tag{C.55}$$

$$= \frac{P(\mathbf{y}; \mathbf{w}_{\overline{\mathbf{y}}} \mid \mathbf{w}_{\mathbf{y}}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}}))}{P(\mathbf{w}_{\overline{\mathbf{y}}} \mid \mathbf{w}_{\mathbf{y}}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}}))} \tag{C.56}$$

$$= \frac{P(\mathbf{y} \mid \mathbf{w}_{\mathbf{y}}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}}))P(\mathbf{w}_{\overline{\mathbf{y}}} \mid \mathbf{y}; \mathbf{w}_{\mathbf{y}}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}}))}{P(\mathbf{w}_{\overline{\mathbf{y}}} \mid \mathbf{w}_{\mathbf{y}}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}}))} \tag{C.57}$$

$$= P(\mathbf{y} \mid \mathbf{w}_{\mathbf{y}}; \mathbf{x}; \mathbf{w}_{\overline{\mathbf{y}}} = do(\mathbf{w}_{\overline{\mathbf{y}}})). \tag{C.58}$$

The second factor in the numerator and the denominator in eq. (C.57) are equal to 1 because

of the intervention $\mathbf{W}_{\bar{\mathbf{y}}} = do(\mathbf{w}_{\bar{\mathbf{y}}})$.

From [10] we have that for any disjoint sets $\mathbf{S}, \mathbf{T} \subseteq \mathbf{V}$, $P(\mathbf{s} \mid do(\mathbf{t})) = P_{d\text{-}sep}[\mathbf{D}]$, where $\mathbf{D} = An(\mathbf{S})$ in $G[\mathbf{V} \setminus \mathbf{T}]$. It is easy to verify that the ancestors of $\mathbf{Y} \cup \mathbf{W}_{\mathbf{y}}$ in $G_{\overline{\mathbf{X}}}$, $\mathbf{A}$, are the same as in $G_{\mathbf{X}}[\mathbf{V} \setminus \mathbf{W}_{\bar{\mathbf{y}}}]$, then keeping $\mathbf{x}$ fixed we could write

$$P(\mathbf{y}, \mathbf{w}_{\mathbf{y}}, \mathbf{x}, \mathbf{w}_{\bar{\mathbf{y}}} = do(\mathbf{w}_{\bar{\mathbf{y}}})) = \sum_{\mathbf{a} \in An(\mathbf{y} \cup \mathbf{w}_{\bar{\mathbf{y}}})} Q[\mathbf{A}, \mathbf{x}]; \tag{C.59}$$

and then

$$P(\mathbf{y} \mid \mathbf{w}, \mathbf{x}) = \frac{P_{\mathbf{a} \in An(\mathbf{y} \cup \mathbf{w})} Q[\mathbf{A}, \mathbf{x}]}{P_{\mathbf{a} \in An(\mathbf{w})} Q[\mathbf{A}, \mathbf{x}]}. \tag{C.60}$$

Let $G^{\Delta'}$ be the same as G^{Δ} after removing all edges out of $\mathbf{W}_{\mathbf{y}}$ and any edge out of $\mathbf{Y}$ that is not part of a directed from $\mathbf{Y}$ to $\mathbf{W}_{\mathbf{y}}$. $G^{\Delta'}$ and G^{Δ} have the same c-component structure and all variables in $\mathbf{A}$ are still ancestors of $\mathbf{Y} \cup \mathbf{W}$; therefore and by the same reasoning, $Q[\mathbf{A}, \mathbf{x}]$ is not transportable from $\langle G^{\Delta'}, Z \rangle$ either. Without loss of generality let $M^{(i)} = \{M^{i,k}\}_{k \in \mathcal{K}}$, $i = 1, 2$, be sets of models witnessing the non-transportability of $Q[\mathbf{A}, \mathbf{x}]$ from $\langle G^{\Delta'}, Z \rangle$. Since every variable in $\mathbf{A} \cap (\mathbf{Y} \cup \mathbf{W}_{\mathbf{y}})$ has a children in $\mathbf{A}$, we can apply lemma 12 in a topological order over $\mathbf{A} \cap (\mathbf{Y} \cup \mathbf{W}_{\mathbf{y}})$ and conclude that $Q[\mathbf{A}, \mathbf{x}]$ transportable if and only if $P(\mathbf{y}, \mathbf{w}_{\mathbf{y}}, \mathbf{x}, \mathbf{w}_{\bar{\mathbf{y}}} = do(\mathbf{w}_{\bar{\mathbf{y}}})) = P_{\mathbf{a} \in An(\mathbf{y} \cup \mathbf{w}_{\mathbf{y}})} Q[\mathbf{A}, \mathbf{x}]$ is transportable.

For simplicity, let

$$\gamma(\mathbf{y}, \mathbf{w}) = P(\mathbf{y}, \mathbf{w}_{\mathbf{y}}, \mathbf{x}, \mathbf{w}_{\bar{\mathbf{y}}} = do(\mathbf{w}_{\bar{\mathbf{y}}})) = \sum_{\mathbf{a} \in An(\mathbf{y} \cup \mathbf{w}_{\bar{\mathbf{y}}})} Q[\mathbf{A}, \mathbf{x}]; \tag{C.61}$$

then

$$P(\mathbf{y} \mid \mathbf{w}, \mathbf{x}) = \frac{\gamma(\mathbf{y}, \mathbf{w})}{\sum_{\mathbf{y}} \gamma(\mathbf{y}, \mathbf{w})}. \tag{C.62}$$

If $(\mathbf{w}) = \sum_{\mathbf{y}} (\mathbf{y}; \mathbf{w})$ is transportable then $P(\mathbf{y} \mid \mathbf{w}; \mathbf{x})$ must be non-transportable, else

$$(\mathbf{y}; \mathbf{w}) = P(\mathbf{y} \mid \mathbf{w}; \mathbf{x})\ (\mathbf{w}); \tag{C.63}$$

contradicting the assumption that the $(\mathbf{y}; \mathbf{w})$ is not transportable. Therefore, we can further assume $(\mathbf{w})$ is not transportable, for the rest of the argument.

Let $(\mathbf{y}'; \mathbf{w}')$ be an assignment such that $(\mathbf{y}'; \mathbf{w}')$ differs among $M^{(1)}$ and $M^{(2)}$, then let

$\mathbf{W}$	$\mathbf{Y}$	$^{(1)}(\mathbf{W}; \mathbf{Y})$	$^{(2)}(\mathbf{W}; \mathbf{Y})$
$\mathbf{w}'$	$\mathbf{y}'$	a	b
$\mathbf{w}'$	$\neq \mathbf{y}'$	c	d
$\neq \mathbf{w}'$	$\mathbf{y}'$	e	f
$\neq \mathbf{w}'$	$\neq \mathbf{y}'$	g	h

Due to the non-transportability of $(\mathbf{y}; \mathbf{w})$ we have $a \neq b$ and without loss of generality we can assume $a > b$. Similarly, due to the non-transportability of $(\mathbf{w})$ we have $a + c \neq b + d$. For $M^{(1)}$ and $M^{(2)}$:

$$^{(1)}(\mathbf{y}' \mid \mathbf{w}') = \frac{a}{a + c} \tag{C.64}$$

$$^{(2)}(\mathbf{y}' \mid \mathbf{w}') = \frac{b}{b + d}; \tag{C.65}$$

These probabilities are equal if and only if $ad = bc$. Hence, if they are not equal we are done because $^{(1)}(\mathbf{y}' \mid \mathbf{w}') \neq {}^{(2)}(\mathbf{y}' \mid \mathbf{w}')$. If they are equal, let $W \in \mathbf{W}_{\mathbf{y}}$ be such that there exists a path $\bar{p}$ between it and $Y \in \mathbf{Y}$ in $G^{\Delta'}$ that does not contain any $\mathbf{W} \cap \{W\}$. Such $\bar{p}$ exists because by assumption every element in $\mathbf{W}_{\mathbf{y}}$ is connected to some element of $\mathbf{Y}$ in $G_{\mathbf{x}}[\mathbf{D}]_{\underline{\mathbf{W}}}$ (which is a subgraph of $G^{\Delta'}$), so W could be just the closest to some Y.

Add a bit to every variable in $\bar{p}$ and denote them with subscript p. Define independent functions for the bits which we will parametrize later.

We define two new models $M^{(1)\prime}$ and $M^{(2)\prime}$, based on $M^{(1)}$ and $M^{(2)}$. For every variable in $\bar{p}$ except for W and Y, append the corresponding extra bit defined in $\bar{p}$ with the original variables in $M^{(1)}$ and $M^{(2)}$. Rename W and Y as W_p; Y_p and make them unobservable, then define W in the new models with the functions:

$$f'_w = \begin{cases} w' & \text{if } W_p = 1 \\ W_p & \text{otherwise;} \end{cases} \tag{C.66}$$

where W_p is unobservable too and w' is the assignment to W consistent with $\mathbf{w}'$ for which the query disagrees in $M^{(1)}$ and $M^{(2)}$.

Analogously, define

$$f'_y = \begin{cases} y' & \text{if } Y_p = 1 \\ Y_p & \text{otherwise:} \end{cases} \tag{C.67}$$

The path $\bar{p}$ must have a common ancestor to W and Y. Such ancestor could be observable or unobservable. That is, either there exists $Z \in \bar{p} \setminus An(W) \setminus An(Y)$ (possibly Y itself) or there exists $Z_1; Z_2 \in \bar{p}$ with $Z_1 \in An(W); Z_2 \in An(Y)$ and there is an edge $Z_1 \to Z_2$ in $\bar{p}$. For the parametrization of the extra bits in $\bar{p}$ define a new unobservable U and let $Z_p = U$ if the common ancestor is observable or let U be the unobservable parent of Z_{1p} and Z_{2p} in the second. Notice that Z_1 and Z_2 may be equal to W and Y themselves.

Using lemma 18 we will parametrize $\bar{p}$ such that

W_0	Y_0	$P(W_0; Y_0)$
1	1	$\frac{1}{2}$
1	0	$\frac{1}{2}(1$ $)$
0	1	$\frac{1}{2}(1$ $)$
0	0	$\frac{1}{2}$

for some 2 $(0; 1)$ that we will pick later.

Claim 1 (Disagreement on the query)**.** *$M^{(1)'}$ and $M^{(2)'}$ disagree on the query for any such that c $d \in [(a + c + 1)h$ $(b + d + 1)g](1$ $)$.*

Proof. For $M^{(1)'}$ and $M^{(2)'}$ we have

$$^{0}(\mathbf{w}^{0}; \mathbf{y}^{0}) = \sum_{w_p; y_p; \mathbf{w}; \mathbf{y}} {}^{0}(\mathbf{w}^{0}; \mathbf{y}^{0}; \mathbf{w}; \mathbf{y} \mid w_p; y_p) P(w_p; y_p): \tag{C.68}$$

Going over each possible combination of W_p and Y_p first we get

$$^{0}(\mathbf{w}^{0}; \mathbf{y}^{0}) = (\mathbf{W} = \mathbf{w}^{0}; \mathbf{Y} = \mathbf{y}^{0}) P(W_p = 0; Y_p = 0) \tag{C.69}$$

$$+ (\mathbf{Y} = \mathbf{y}^{0}) P(W_p = 1; Y_p = 0) \tag{C.70}$$

$$+ (\mathbf{W} = \mathbf{w}^{0}) P(W_p = 0; Y_p = 1) \tag{C.71}$$

$$+ P(W_p = 1; Y_p = 1): \tag{C.72}$$

Similarly,

$$^{0}(\mathbf{w}^{0}) = (\mathbf{W} = \mathbf{w}^{0}) P(W_p = 0) + P(W_p = 1): \tag{C.73}$$

For $M^{(1)'}$

$$
\begin{aligned}
{}^{(1)'}(\mathbf{y}' \mid \mathbf{w}') &= \frac{\frac{1}{2}a \quad + \frac{1}{2}(a+e)(1 \quad) + \frac{1}{2}(a+c)(1 \quad) + \frac{1}{2}}{\frac{1}{2}(a+c) + \frac{1}{2}} && \text{(C.74)}\\
&= \frac{a \quad + (2a+c+e)(1 \quad) +}{a+c+1} && \text{(C.75)}\\
&= \frac{a \quad (a+c+e)(\quad 1) +}{a+c+1} && \text{(C.76)}\\
&= \frac{a + (a+c+e)(1 \quad) \quad (1 \quad) + 1}{a+c+1} && \text{(C.77)}\\
&= \frac{a \quad (1 \quad (a+c+e))(1 \quad) + 1}{a+c+1} && \text{(C.78)}\\
&= \frac{a \quad g(1 \quad) + 1}{a+c+1} && \text{(C.79)}
\end{aligned}
$$

Analogously for $M^{(2)'}$:

$$
{}^{(2)'}(\mathbf{y}' \mid \mathbf{w}') = \frac{b \quad h(1 \quad) + 1}{b+d+1} \qquad \text{(C.80)}
$$

Those two are equal if and only if

$$
\begin{aligned}
& ab \quad bg(1 \quad) + b + ad \quad dg(1 \quad) + d + a \quad g(1 \quad) + 1 \\
&= ab \quad ah(1 \quad) + a + bc \quad ch(1 \quad) + c + b \quad h(1 \quad) + 1 && \text{(C.81)}\\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad ( ) \\
& \qquad bg(1 \quad) + ad \quad dg(1 \quad) + d + \quad g(1 \quad) \\
&= \quad ah(1 \quad) + bc \quad ch(1 \quad) + c + \quad h(1 \quad) && \text{(C.82)}
\end{aligned}
$$

Recall that we have $ad = bc$, which also implies that $c \neq d$, else $a = b$ which is a contradiction. Then, the condition for equality can be further simplified as

$$
c \quad d = [(a+c+1)h \quad (b+d+1)g](1 \quad): \qquad \text{(C.83)}
$$

The left hand side is non-zero, all $a; b; c; d; g$ and h are fixed, and is a free parameter.

Therefore, as long as we pick a $\mathbf{x}$ such that the equality doesn't hold, we get that $(\mathbf{y}' \mid \mathbf{w}') = P(\mathbf{y}' \mid \mathbf{w}'; \mathbf{x})$ does not match in $M^{(1)'}$ and $M^{(2)'}$. □

Claim 2 (Agreement on any agreed distribution)**.** *Let* $\mathbf{Z} \subseteq \mathbf{V}$ *be any subset of observable variables and let* $\mathbf{z} \in \mathcal{Z}^k \in \mathcal{Z}$. *If* $M^{(1)}$ *and* $M^{(2)}$ *agree on* $P^k(\mathbf{V}; \mathbf{z})$, *then* $M^{(1)'}$ *and* $M^{(2)'}$ *also agree on* $P^k(\mathbf{V}; \mathbf{z})$.

Proof. For simplicity we omit the superscript k for the domain, which is fixed with $\mathbf{z}$. The superscript (1) and (2) indicate to which of the sets of models under consideration the expression refers to.

Let $\mathbf{C}_1, \mathbf{C}_2, \ldots$ be the C-components of $G_{\overline{\mathbf{z}}}$. By assumption we have $Q^{(1)}[\mathbf{V}; \mathbf{z}] = Q^{(2)}[\mathbf{V}; \mathbf{z}]$, and since any $Q[\mathbf{C}_j; \mathbf{z}]$ is identifiable from $Q[\mathbf{V}; \mathbf{z}]$, we have $Q^{(1)}[\mathbf{C}_j; \mathbf{z}] = Q^{(2)}[\mathbf{C}_j; \mathbf{z}]$ for any $\mathbf{C}_j$.

$M^{(k)'}$ is identical to $M^{(k)}$, $k = 1, 2$, except for the functions of the observables in the path $\overline{p}$. For any variable T not in $\overline{p}$, but with a parent on it, the function f_T remains the same and it simply ignores the extra bit that its parent has in $M^{(k)'}$.

Let $\mathbf{C}_j$ be a C-component containing some set of variables $\mathbf{R}$ in $\overline{p}$ different to W and Y (the endpoints of $\overline{p}$). First, by definition

$$Q^{(k)'}[\mathbf{C}_j; \mathbf{z}](\mathbf{v}) = \sum_{\mathbf{u}(\mathbf{C}_j)} \prod_{V_i \in \mathbf{C}_j} P^{(k)'}(v_i \mid \mathbf{pa}_i, \mathbf{u}_i; \mathbf{x}) P^{(k)'}(\mathbf{u}(\mathbf{C}_j)). \tag{C.84}$$

For any $S \in \overline{p}$, $R \in \mathbf{R}$, W and Y that could be in $\mathbf{C}_j$ in $M^{k'}$, their corresponding factors in

the previous expression can be re-written in terms of probabilities of M^k, as follows:

$$P^{(k)'}(s \mid \mathbf{pa}_s; u_s; \ \mathbf{x}) = P^{(k)}(s \mid \mathbf{pa}_s; u_s; \ \mathbf{x}); \tag{C.85}$$

$$P^{(k)'}(r \mid \mathbf{pa}_r; u_r; \ \mathbf{x}) = P^{(k)}(r \mid \mathbf{pa}_r; u_r; \ \mathbf{x}) P(r_p \mid (\mathbf{pa}_r)_p); \tag{C.86}$$

$$\begin{aligned} P^{(k)'}(y \mid \mathbf{pa}_y; u_y; \ \mathbf{x}) = {} & P^{(k)}(y \mid \mathbf{pa}_y; u_y; \ \mathbf{x}) P(Y_p = 0 \mid (\mathbf{pa}_y)_p) \\ & + 1[y = y']P(Y_p = 1 \mid (\mathbf{pa}_y)_p)); \end{aligned} \tag{C.87}$$

$$\begin{aligned} P^{(k)'}(w \mid \mathbf{pa}_w; u_w; \ \mathbf{x}) = {} & P^{(k)}(w \mid \mathbf{pa}_w; u_w; \ \mathbf{x}) P(W_p = 0 \mid (\mathbf{pa}_w)_p) \\ & + 1[w = w']P(W_p = 1 \mid (\mathbf{pa}_w)_p)): \end{aligned} \tag{C.88}$$

It follows that

$$\begin{aligned} Q^{(k)'}[\mathbf{C}_j; \ \mathbf{z}](\mathbf{v}) = {} & \Bigg(\prod_{R \in \mathbf{R}} P(r_p \mid (\mathbf{pa}_r)_p) \Bigg) \\ & Q^{(k)}[\mathbf{C}_j; \ \mathbf{z}](\mathbf{v}) P(Y_p = 0 \mid (\mathbf{pa}_y)_p) P(W_p = 0 \mid (\mathbf{pa}_w)_p) \\ & + Q^{(k)}[\mathbf{C}_j \cap \{Y\}; \ \mathbf{z}](\mathbf{v}) P(Y_p = 1 \mid (\mathbf{pa}_y)_p) P(W_p = 0 \mid (\mathbf{pa}_w)_p) 1[y = y'] \\ & + Q^{(k)}[\mathbf{C}_j \cap \{W\}; \ \mathbf{z}](\mathbf{v}) P(Y_p = 0 \mid (\mathbf{pa}_y)_p) P(W_p = 1 \mid (\mathbf{pa}_w)_p) 1[w = w'] \\ & + Q^{(k)}[\mathbf{C}_j \cap \{W; Y\}; \ \mathbf{z}](\mathbf{v}) P(Y_p = 1 \mid (\mathbf{pa}_y)_p) P(W_p = 1 \mid (\mathbf{pa}_w)_p) 1[w = w'; y = y'] \\ & : \end{aligned} \tag{C.89}$$

Since W and Y have no descendants in G,

$$Q^{(k)}[\mathbf{C}_j \cap \{W; Y\}; \ \mathbf{z}](\mathbf{v}) = \sum_{w;y} Q^{(k)}[\mathbf{C}_j; \ \mathbf{z}](\mathbf{v}) \tag{C.90}$$

$$Q^{(k)}[\mathbf{C}_j \cap \{W\}; \ \mathbf{z}](\mathbf{v}) = \sum_{w} Q^{(k)}[\mathbf{C}_j; \ \mathbf{z}](\mathbf{v}) \tag{C.91}$$

$$Q^{(k)}[\mathbf{C}_j \cap \{Y\}; \ \mathbf{z}](\mathbf{v}) = \sum_{y} Q^{(k)}[\mathbf{C}_j; \ \mathbf{z}](\mathbf{v}); \tag{C.92}$$

all match between $M^{(1)}$ and $M^{(2)}$. Consequently, every C-factor in the right-hand side

of (C.89) is the same in those models, and since every other term is also the same in both $M^{(1)\prime}$ and $M^{(2)\prime}$, we conclude that $Q^{(1)\prime}[\mathbf{C}_j;\ \mathbf{z}](\mathbf{v}) = Q^{(2)\prime}[\mathbf{C}_j;\ \mathbf{z}](\mathbf{v})$, which in turn implies our claim since

$$P^{(k)\prime}(\mathbf{v};\ \mathbf{z}) = \prod_j Q^{(k)\prime}[\mathbf{C}_j;\ \mathbf{z}](\mathbf{v}). \tag{C.93}$$

□

In summary, $M^{(1)\prime}$ and $M^{(2)\prime}$ induce G^{Δ} and matching Z, yet they differ on the value for $P(\mathbf{y}' \mid \mathbf{w}';\ \mathbf{x})$, proving the non-transportability of the query. □

C.5 General Causal Data Fusion Tasks

Theorem 22 (C-INFER soundness)**.** *Given a causal inference task with signature $I = \langle Q; \mathsf{P}; G \rangle$, where the query is a conditional associational or interventional query, each input $P \in \mathsf{P}$ is an observational/interventional and partially observed/selection biased distribution, and G is one or more causal diagrams or selection diagrams; then Q can be evaluated as a function of P if* C-INFER *finds a mapping using the cftree-operators and G. Moreover, the process takes $O(n^2(n+m)p)$ time, where $n = |\mathbf{V}|$, m is the number of edges in G and $p = |\mathsf{P}|$ if there is at most one partially observed distribution in P.*

Proof. **Soundness.** This part follows from the soundness of the individual cftree operators.

Efficiency. As long as there is at most one distribution measuring a subset of $\mathbf{V}$, the *q tree* T_Q generated by GENQUERYTREE will have $O(n)$ target nodes, hence the overall running time is the same as for other tasks, that is, $O(n^2(n+m)p)$ (see the proof for theorem 6 for instance). □

Appendix D: Soft Interventions and Sigma Calculus

D.1 Relationship between Soft and Atomic Interventions

Proof. We can start by summing over the $\mathbf{D} \cap \mathbf{Y}$:

$$P(\mathbf{y};\ \mathbf{x} = \mathbf{x}) = \sum_{\mathbf{d} \cap \mathbf{y}} P(\mathbf{d};\ \mathbf{x} = \mathbf{x}). \tag{D.1}$$

Let $V_1 < V_2 < \cdots$ be a topological order of the variables in $\mathbf{D}$ in the graph $G_{\mathbf{x}}$ and let $\mathbf{V}^{<i}$ be the set of variables in $\mathbf{D}$ that comes before V_i in the order. Then we can write

$$P(\mathbf{y};\ \mathbf{x} = \mathbf{x}) = \sum_{\mathbf{d} \cap \mathbf{y}} \prod_{V_i \in \mathbf{D}} P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x} = \mathbf{x}). \tag{D.2}$$

For every $V_i \in \mathbf{X}$ let $\mathbf{X}^{<i} = \mathbf{X} \setminus \mathbf{V}^{<i}$ and $\mathbf{X}^{>i} = \mathbf{X} \cap \mathbf{V}^{<i}$. First we will use rule 3 of σ-calculus to exchange $\mathbf{x}$ with an atomic intervention for the variables in $\mathbf{X}$ that go after V_i. This is allowed by $(V_i \perp\!\!\!\perp \mathbf{X}^{>i} \mid \mathbf{V}^{<i})$ in both $G_{\mathbf{x}\overline{\mathbf{X}^{>i}}}$ and $G_{\mathbf{x}^{<i}(\mathbf{x}^{>i} = \mathbf{x}^{>i})\overline{\mathbf{X}^{>i}}}$.

$$P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x} = \mathbf{x}) = P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x}^{<i} = \mathbf{x}^{<i};\ \mathbf{x}^{>i} = \mathbf{x}^{>i}); \tag{D.3}$$

In other words, V_i is not affected by interventions on variables that come before as long as we do not condition on their descendants.

Next, we will do the same for variables in $\mathbf{X}^{>i}$. The difference is that those are observed because they belong to $\mathbf{V}^{<i}$. The separation statement $(V_i \perp\!\!\!\perp \mathbf{X}^{<i} \mid \mathbf{V}^{<i} \cap \mathbf{X})$ holds in $G_{(\mathbf{x}^{>i} = \mathbf{x}^{>i})\,\mathbf{x}^{<i}\underline{\mathbf{X}^{<i}}}$ and $G_{(\mathbf{x}^{>i} = \mathbf{x}^{>i})(\mathbf{x}^{<i} = \mathbf{x}^{<i})\underline{\mathbf{X}^{<i}}}$. To see why consider any $X \in \mathbf{X}^{<i}$ that may be connected to V_i by an active path in those graphs. The path must have arrows into X and the arrow is not bidirected because under intervention x no such edge appears in

the graph. Then, the arrow must be direct and the observable variable at the tail is in $\mathbf{V}^{<i}$, so the path is blocked. By those separations and rule 2 it follows:

$$P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x} = \mathbf{x}) = P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x}^{<i}=\mathbf{x}^{<i};\ \mathbf{x}^{>i}=\mathbf{x}^{>i}) \tag{D.4}$$

$$= P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x}=\mathbf{x}). \tag{D.5}$$

At this point all interventions are atomic. Using the definition of conditional probability:

$$P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x}=\mathbf{x}) = \frac{P(v_i \mid \mathbf{v}^{<i} \cap \mathbf{x}^{<i};\ \mathbf{x}=\mathbf{x})P(\mathbf{x}^{<i} \mid v_i; \mathbf{v}^{<i} \cap \mathbf{x}^{<i};\ \mathbf{x}=\mathbf{x})}{P(\mathbf{x}^{<i} \mid \mathbf{v}^{<i} \cap \mathbf{x}^{<i};\ \mathbf{x}=\mathbf{x})}. \tag{D.6}$$

The second factor in the numerator and the denominator are equal to 1 because under $\mathbf{x}=\mathbf{x}$ the intervened variables always take the specified values, in summary

$$P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x} = \mathbf{x}) = P(v_i \mid \mathbf{v}^{<i} \cap \mathbf{x}^{<i};\ \mathbf{x}=\mathbf{x}). \tag{D.7}$$

Then we can write the original effect as

$$P(\mathbf{y};\ \mathbf{x} = \mathbf{x}) = \sum_{\mathbf{d} \cap \mathbf{y}} \prod_{V_i \in \mathbf{D} \cap \mathbf{X}} P(v_i \mid \mathbf{v}^{<i} \cap \mathbf{x}^{<i};\ \mathbf{x}=\mathbf{x}) \prod_{V_i \in \mathbf{D} \setminus \mathbf{X}} P(v_i \mid \mathbf{v}^{<i};\ \mathbf{x} = \mathbf{x}). \tag{D.8}$$

The first product can be simply written as $P(\mathbf{d} \cap \mathbf{x};\ \mathbf{x} = \mathbf{x})$ by virtue of the product/chain rule of probabilities. In the second product, for every $V_i \in \mathbf{X}$ we have $(V_i \perp\!\!\!\perp \mathbf{V}^{<i} \cap \mathbf{Pa}_i \mid \mathbf{Pa}_i)$. Finally we get

$$P(\mathbf{y};\ \mathbf{x} = \mathbf{x}) = \sum_{\mathbf{d} \cap \mathbf{y}} P(\mathbf{d} \cap \mathbf{x};\ \mathbf{x}=\mathbf{x}) \prod_{X \in \mathbf{D} \setminus \mathbf{X}} P(x \mid \mathbf{pa}_x;\ \mathbf{x} = \mathbf{x}); \tag{D.9}$$

which corresponds to eq. (4.22).

Since every variable in $\mathbf{X}$ has its own c-component in $G_{\underline{\mathbf{X}}}$, by -operator we have

$$Q[\mathbf{D};\ _{\mathbf{X}}] = Q[\mathbf{D} \cap \mathbf{X};\ _{\mathbf{X}}]Q[\mathbf{X} \setminus \mathbf{D};\ _{\mathbf{X}}]. \tag{D.10}$$

Then by -operator,

$$Q[\mathbf{D} \cap \mathbf{X};\ _{\mathbf{X}}] = Q[\mathbf{D} \cap \mathbf{X}] = Q[\mathbf{D} \cap \mathbf{X};\ _{\mathbf{X}} = \mathbf{x}]. \tag{D.11}$$

If $Q[\mathbf{D} \cap \mathbf{X};\ _{\mathbf{X}} = \mathbf{x}]$ is not identifiable, $Q[\mathbf{D};\ _{\mathbf{X}}]$ is also not identifiable, because whenever the latter is identifiable the former can be computed via eq. (D.10). Moreover, $Q[\mathbf{D} \cap \mathbf{X};\ _{\mathbf{X}} = \mathbf{x}]$ is the same as $P(\mathbf{d} \cap \mathbf{x};\ _{\mathbf{X}} = \mathbf{x})$.

Finally, the necessity of $Q[\mathbf{D};\ _{\mathbf{X}}]$ for $P(\mathbf{y};\ _{\mathbf{X}})$ follows from applying lemma 12 in a topological order over the elements of $\mathbf{D} \cap \mathbf{Y}$. □

D.2 Soundness of -calculus

Theorem 5. *[Inference Rules — -calculus] Let G be a causal diagram compatible with an SCM M, with endogenous variables $\mathbf{V}$. For any disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, two disjoint subsets $\mathbf{T}, \mathbf{W} \subseteq \mathbf{V} \cap (\mathbf{Z} \cup \mathbf{Y})$ (i.e., possibly including $\mathbf{X}$), the following rules are valid for any intervention strategies $_{\mathbf{X}}$, $_{\mathbf{Z}}$, and $^{()}_{\mathbf{Z}}$ such that $G_{\underline{\mathbf{X}}\overline{\mathbf{Z}}}$ and $G_{\underline{\mathbf{X}}\overline{\mathbf{Z}(\mathbf{W})}}$ have no cycles:*

Rule 1 (Insertion/Deletion of observations):

$$P(\mathbf{y} \mid \mathbf{w}, \mathbf{t};\ _{\mathbf{X}}) = P(\mathbf{y} \mid \mathbf{w};\ _{\mathbf{X}}) \quad \text{if } (\mathbf{T} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}) \text{ in } G_{\underline{\mathbf{X}}}. \tag{4.23}$$

Rule 2 (Change of regimes under observation):

$$P(\mathbf{y} \mid \mathbf{z}, \mathbf{w};\ _{\mathbf{x}},\ _{\mathbf{z}}) = P(\mathbf{y} \mid \mathbf{z}, \mathbf{w};\ _{\mathbf{x}},\ ^{()}_{\mathbf{z}}) \quad \text{if } (\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}) \text{ in } G_{\underline{\mathbf{X}}\,\overline{\mathbf{Z}}} \text{ and } G_{\underline{\mathbf{X}}\,\overline{\mathbf{Z}(\mathbf{W})}}. \tag{4.24}$$

Rule 3 (Change of regimes without observation):

$$P(\mathbf{y} \mid \mathbf{w}; \hat{\mathbf{x}}, \hat{\mathbf{z}}) = P(\mathbf{y} \mid \mathbf{w}; \hat{\mathbf{x}}, \hat{\mathbf{z}}') \quad \text{if } (\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}) \text{ in } G_{\hat{\mathbf{x}}\overline{\mathbf{z(w)}}} \text{ and } G_{\hat{\mathbf{x}}\hat{\mathbf{z}}'\overline{\mathbf{z(w)}}}; \quad (4.25)$$

where $\mathbf{Z(W)}$ $\mathbf{Z}$ *is the set of elements in* $\mathbf{Z}$ *that are not ancestors of* $\mathbf{W}$ *in* $G_{\hat{\mathbf{x}}}$.

Proof. **Rule 1** Both sides of the expression refer to the same model M_x and the corresponding causal diagram G_x. So the condition licenses the equality by application of the d-separation criterion in the context of this pair.

To prove the next two rules, we will consider a new SCM M with observable variables $\mathbf{V} = \mathbf{V} \cup \{\mathbf{z}, \mathbf{z} = f_Z(g_{Z2}\mathbf{z}\}$, that is a new node for each variable affected by interventions on variables in $\mathbf{Z}$. M has a set of unobservables $\mathbf{U}$ $\mathbf{U}$, and the distribution $P(\mathbf{u})$ is the same. Further, M has a set of functions F such that $f_{V_i} = f_{V_i}$ for $V_i \in \mathbf{V} \cap (\mathbf{X} \cup \mathbf{Z})$, for $X \in \mathbf{X}$ let $f_X = f_{X_{\hat{\mathbf{x}}}}$ (the function for X in the model $M_{\hat{\mathbf{x}}}$). Finally, for $f_Z; Z \in \mathbf{Z}$ let

$$f_Z = \begin{cases} f_{Z_{\hat{\mathbf{z}}'}} & \text{if } \mathbf{z} = 0 \\ f_Z & \text{if } \mathbf{z} = 1 \end{cases}; \quad (D.12)$$

where f_Z is the function of Z in the model $M_{\hat{\mathbf{z}}}$ and $f_{Z_{\hat{\mathbf{z}}'}}$ the same function in $M_{\hat{\mathbf{z}}'}$.

The model M induces a graph G where $\mathbf{pa}_i$ for any $V_i \in \mathbf{V} \cap \mathbf{Z}$ is the same in as in $G_{\hat{\mathbf{x}}}$, while $\mathbf{pa}_i; V_i \in \mathbf{Z}$ is the union of the parents of Z in $G_{\hat{\mathbf{z}}'}$ and $G_{\hat{\mathbf{z}}}$.

Let P denote the probability distribution induced by M. We have that $P(\mathbf{v} \mid \mathbf{z} = 1)$ is exactly the same as $P(\mathbf{v}; \hat{\mathbf{x}}, \hat{\mathbf{z}})$ while $P(\mathbf{v} \mid \mathbf{z} = 0)$ behaves as $P(\mathbf{v}; \hat{\mathbf{x}}, \hat{\mathbf{z}}')$. It follows that for any pair of disjoint sets $\mathbf{A}; \mathbf{B} \subseteq \mathbf{V}$:

$$P(\mathbf{a} \mid \mathbf{b}; \mathbf{z} = 1) = P(\mathbf{a} \mid \mathbf{b}; \hat{\mathbf{x}}, \hat{\mathbf{z}}); \text{ and} \quad (D.13)$$

$$P(\mathbf{a} \mid \mathbf{b}; \mathbf{z} = 0) = P(\mathbf{a} \mid \mathbf{b}; \hat{\mathbf{x}}, \hat{\mathbf{z}}'): \quad (D.14)$$

Rule 2 If $(\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W}, \mathbf{Z})$ in G it follows

$$P(\mathbf{y} \mid \mathbf{z}, \mathbf{w}, \mathbf{x}, \mathbf{z}) = P(\mathbf{y} \mid \mathbf{z}, \mathbf{w}, \mathbf{z} = 1) \tag{D.15}$$

$$= P(\mathbf{y} \mid \mathbf{z}, \mathbf{w}, \mathbf{z} = 0) \tag{D.16}$$

$$= P(\mathbf{y} \mid \mathbf{z}, \mathbf{w}, \mathbf{x}, \hat{\mathbf{z}}). \tag{D.17}$$

If this independence does not hold, there exists a path from Z to some $Y \in \mathbf{Y}$ in G that is d-connected given $\mathbf{W} \cup \mathbf{Z}$. Let p (without loss of generality) be such a path with no node in $\mathbf{Z}$ other than Z in it. The path p must start with $Z \rightarrow Z$ A, for some variable A, else it is blocked by conditioning on Z. The edge $(Z \rightarrow A)$ is either present in $G_{\mathbf{x}\,\mathbf{z}\underline{\mathbf{Z}}}$ or in $G_{\mathbf{x}\,\hat{\mathbf{z}}\underline{\mathbf{Z}}}$, which implies the portion of p from Z to Y is present in one of those graphs and d-connected given $\mathbf{W}$, which leads to a contradiction to the conditions in the rule.

Rule 3 If $(\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{W})$ in G it follows

$$P(\mathbf{y} \mid \mathbf{w}, \mathbf{x}, \mathbf{z}) = P(\mathbf{y} \mid \mathbf{w}, \mathbf{z} = 1) \tag{D.18}$$

$$= P(\mathbf{y} \mid \mathbf{w}, \mathbf{z} = 0) \tag{D.19}$$

$$= P(\mathbf{y} \mid \mathbf{w}, \mathbf{x}, \hat{\mathbf{z}}). \tag{D.20}$$

If this independence does not hold, there exists a path from Z to some $Y \in \mathbf{Y}$ in G that is d-connected given $\mathbf{W}$. Let p (without loss of generality) be such a path with no node in $\mathbf{Z}$ other than Z in it. If p starts with $Z \rightarrow Z$ A, Z must have a descendant in $\mathbf{W}$ which implies $Z \in \mathbf{Z}(\mathbf{W})$. Hence the edge $(Z \rightarrow A)$ is in $G_{\mathbf{x}\,\mathbf{z}\overline{\mathbf{Z}(\mathbf{W})}}$ or $G_{\mathbf{x}\,\hat{\mathbf{z}}\overline{\mathbf{Z}(\mathbf{W})}}$. If p starts with $Z \rightarrow Z \rightarrow A$, $(Z \rightarrow A)$ is also in one of those graphs.

Then, the portion of p from Z to Y exists either in $G_{\mathbf{x}\,\mathbf{z}\overline{\mathbf{Z}(\mathbf{W})}}$ or $G_{\mathbf{x}\,\hat{\mathbf{z}}\overline{\mathbf{Z}(\mathbf{W})}}$ and is d-connected given $\mathbf{W}$, a contradiction to at least one of the independences in the rule.

□

D.3 Completeness of -calculus for -TR

To prove the completeness of -calculus we will show the relationship between causal effects and c-factors and then how every cftree operator used for -TR follows from -calculus.

First we will show that Lemma 11 follows from -calculus. For simplicity, for any $\mathbf{C}; \mathbf{X} \quad \mathbf{V}$ and intervention $_{\mathbf{X}}$ we will write

$$P(\mathbf{c};\ _{\mathbf{X}\setminus\mathbf{C}} = _{\mathbf{X}\setminus\mathbf{C}};\ _{\mathbf{V}\cap\mathbf{C}} = (\mathbf{v} \cap \mathbf{c})) \tag{D.21}$$

simply as

$$P(\mathbf{c};\ _{\mathbf{X}(\mathbf{C})}); \tag{D.22}$$

and $Q[\mathbf{C};\ _{\mathbf{X}} = _{\mathbf{X}}]$ as $Q[\mathbf{C};\ _{\mathbf{X}}]$.

Lemma 19 (C-factor — Causal Effect)**.**

$$Q[\mathbf{C};\ _{\mathbf{X}}] = P(\mathbf{c};\ _{\mathbf{X}(\mathbf{C})}) \tag{D.23}$$

Proof. From the model $M_{\mathbf{X}(\mathbf{C})}$ we have

$$P(\mathbf{v};\ _{\mathbf{X}(\mathbf{C})}) = \sum_{\mathbf{u}} \prod_{i} P(v_i \mid \mathbf{pa}_i; \mathbf{u}_i;\ _{\mathbf{X}(\mathbf{C})}) P(\mathbf{u}). \tag{D.24}$$

Summing both sides over $\mathbf{V} \cap \mathbf{C}$

$$P(\mathbf{c};\ _{\mathbf{X}(\mathbf{C})}) = \sum_{\mathbf{u}} \sum_{an(\mathbf{c})\cap\mathbf{c}} \prod_{\{i \mid V_i \in An(\mathbf{C})\}} P(v_i \mid \mathbf{pa}_i; \mathbf{u}_i;\ _{\mathbf{X}(\mathbf{C})}) P(\mathbf{u}). \tag{D.25}$$

With $_{\mathbf{X}(\mathbf{C})}$ any variable in $V_i \in An(\mathbf{C}) \cap \mathbf{C}$ has been fixed to a constant, so each such factor

$P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \mathbf{x}_{(C)})$ is 1 when the index of the sum over $An(C) \cap \mathbf{C}$ is consistent with $\mathbf{v}$ and 0 otherwise, then

$$P(\mathbf{c}, \mathbf{x}_{(C)}) = \sum_{\mathbf{u}} \prod_{\{i \mid V_i \in \mathbf{C}\}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \mathbf{x}_{(C)}) P(\mathbf{u}). \tag{D.26}$$

Moreover, any $\mathbf{U}$ that does not appear in $\mathbf{U}(\mathbf{C})$ can be summed out, leaving

$$P(\mathbf{c}, \mathbf{x}_{(C)}) = \sum_{\mathbf{u}(\mathbf{c})} \prod_{\{i \mid V_i \in \mathbf{C}\}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \mathbf{x}_{(C)}) P(\mathbf{u}(\mathbf{c})). \tag{D.27}$$

For any $V_i \in \mathbf{C}$ the factor $P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \mathbf{x}_{(C)}) = 1$ if and only if f_i in $M_{\mathbf{x}_{(C)}}$ evaluates to v_i. The only such f_i affected by $\mathbf{x}_{(C)} = \mathbf{x} = \mathbf{x}, \mathbf{v}_{C} = (\mathbf{v} \cap \mathbf{c})$ are those for $V_i \in \mathbf{X} \setminus \mathbf{C}$, and only because of the $\mathbf{x} = \mathbf{x}$ portion, then

$$P(\mathbf{c}, \mathbf{x}_{(C)}) = \sum_{\mathbf{u}(\mathbf{c})} \prod_{\{i \mid V_i \in \mathbf{C}\}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i, \mathbf{x}) P(\mathbf{u}(\mathbf{c})), \tag{D.28}$$

which is exactly $Q[\mathbf{C}, \mathbf{x}_{(C)}]$ by definition. □

Lemma 20 (C-component decomposition — τ-operator)**.** *Let $\mathbf{C}_1, \ldots, \mathbf{C}_l$ be the C-components of $G_{\mathbf{x}[C]}$, let $C_1 < C_2 < C_n$ be any topological order of the variables in $\mathbf{C}$. Then by τ-calculus and probability axioms we have*

$$P(\mathbf{c}, \mathbf{x}_{(C)}) = \prod_{j} P(\mathbf{c}_j, \mathbf{x}_{(C_j)}), \tag{D.29}$$

where each

$$P(\mathbf{c}_j, \mathbf{x}_{(C_j)}) = \prod_{\{C_i \in \mathbf{C}_j\}} P(c_i \mid c_1, \ldots, c_{i-1}, \mathbf{x}_{(C)}). \tag{D.30}$$

Proof.

$$P(\mathbf{c};\ \mathbf{x}_{(\mathbf{C})}) = \prod_i P(c_i \mid c_1,\ldots,c_{i-1};\ \mathbf{x}_{(\mathbf{C})}) \tag{D.31}$$

Let $\mathbf{B}_i = \{C_1,\ldots,C_{i-1}\} \cap \mathbf{C}_j$ (those variables before C_i not in the same C-component as C_i). Similarly, let $\mathbf{D}_i = \{C_{i+1},\ldots,C_l\} \cap \mathbf{C}_j$.

We have $(C_i \perp\!\!\!\perp \mathbf{D}_i \mid C_1,\ldots,C_{i-1})$ in both $G_{\mathbf{x}_{(\mathbf{C})}\overline{\mathbf{D}_i}}$ and $G_{\mathbf{x}_{(\{C_1,\ldots,C_i\}\cap\mathbf{C}_j)}\overline{\mathbf{D}_i}}$ because any relevant path would start going out of a variable in $C' \in \mathbf{D}_i$ and is either blocked by a non-observed collider or would entail that C_i goes after C' in the order, which is a contradiction. Then by rule 3 we can exchange $\mathbf{x}_{(\mathbf{C})}$ with $\mathbf{x}_{(\{C_1,\ldots,C_i\}\cap\mathbf{C}_j)}$.

Next, $(C_i \perp\!\!\!\perp \mathbf{B}_i \mid \{C_1,\ldots,C_{i-1}\} \cap \mathbf{B}_i)$ in both $G_{\mathbf{x}_{(\{C_1,\ldots,C_i\}\cap\mathbf{C}_j)}\underline{\mathbf{B}_i}}$ and $G_{\mathbf{x}_{(\mathbf{C}_j)}\underline{\mathbf{B}_i}}$. Any path violating these separations must have an into C_i and an arrow into some $C' \in \mathbf{B}_i$. Since C_i and C' are not in the same C-component, the path must have at least one directed arrow in it; but the variable at the tail of such arrow is either observed (is in the same C-component as C_i) or has the outgoing arrows removed (it is in $\mathbf{B}_i$), in any case the path is blocked or non-existent. Then, by rule 2 the intervention can be changed to $\mathbf{x}_{(\mathbf{C}_j)}$ and we have

$$P(c_i \mid c_1,\ldots,c_{i-1};\ \mathbf{x}_{(\mathbf{C})}) = P(c_i \mid c_1,\ldots,c_{i-1};\ \mathbf{x}_{(\mathbf{C}_j)}). \tag{D.32}$$

Under intervention $\mathbf{x}_{(\mathbf{C}_j)}$ any variable in $\mathbf{B}_i$ has been fixed to a constant then

$$P(c_i|c_1,\ldots,c_{i-1};\ \mathbf{x}_{(\mathbf{C}_j)}) = \frac{P(c_i|fc_1,\ldots,c_{i-1}\cap\mathbf{b}_i;\ \mathbf{x}_{(\mathbf{C}_j)})P(\mathbf{b}_i|fc_1,\ldots,c_i\cap\mathbf{b}_i;\ \mathbf{x}_{(\mathbf{C}_j)})}{P(\mathbf{b}_i|fc_1,\ldots,c_{i-1}\cap\mathbf{b}_i;\ \mathbf{x}_{(\mathbf{C}_j)})} \tag{D.33}$$

$$= P(c_i|fc_1,\ldots,c_{i-1}\cap\mathbf{b}_i;\ \mathbf{x}_{(\mathbf{C}_j)}); \tag{D.34}$$

because the second factor of the denominator and the denominator are equal to 1. Reorga-

nizing the factors by C-components we get

$$P(\mathbf{c};\ _{\mathbf{x}(\mathbf{C})}) = \prod_{j}\ \prod_{\{i|C_i \subseteq \mathbf{C}_j\}} P(c_i \mid \{c_1,\ldots,c_{i-1}\} \cap \mathbf{b}_i;\ _{\mathbf{x}(\mathbf{C}_j)}) \tag{D.35}$$

$$= \prod_{j} P(\mathbf{c}_j;\ _{\mathbf{x}(\mathbf{C}_j)}); \tag{D.36}$$

which matches Eq. (D.30). Eq. (D.35) together with Eq. (D.34) imply Eq. (D.30). □

Lemma 21 (Marginalization — -operator)**.** *Each step of* IDENTIFY *follows from -calculus.*

Proof. Suppose $\mathbf{A} = An(\mathbf{C})_{G_{[\mathbf{T}]}} = \mathbf{C}$, that is, every ancestor of $\mathbf{C}$ in $G_{[\mathbf{T}]}$ is already in $\mathbf{C}$. First we argue that $(\mathbf{C} \perp\!\!\!\perp \mathbf{T} \cap \mathbf{C})$ in $G_{\mathbf{x}(\mathbf{C})\overline{\mathbf{T}\cap\mathbf{C}}}$ and $G_{\mathbf{x}(\mathbf{T})\overline{\mathbf{T}\cap\mathbf{C}}}$. In those graphs all arrows incoming to variables not in $\mathbf{C}$, including $\mathbf{T} \cap \mathbf{C}$, are cut. Then, any path between some $T \in \mathbf{T} \cap \mathbf{C}$ and some $C \in \mathbf{C}$ must start with an arrow going out from T. If the other end of the edge is not in $\mathbf{C}$ then the path does not exists in the graphs mentioned before. If the edge goes to some variable in $\mathbf{C}$ we have that T is an ancestor of $\mathbf{C}$ in $G_{[\mathbf{T}]}$ which is assumed not to be the case. Then by rule 3 of -calculus:

$$P(\mathbf{c};\ _{\mathbf{x}(\mathbf{C})}) = P(\mathbf{c};\ _{\mathbf{x}(\mathbf{T})}); \tag{D.37}$$

and summing over $\mathbf{T} \cap \mathbf{C}$:

$$P(\mathbf{c};\ _{\mathbf{x}(\mathbf{C})}) = \sum_{\mathbf{t}\cap\mathbf{c}} P(\mathbf{t};\ _{\mathbf{x}(\mathbf{T})}); \tag{D.38}$$

as desired. □

Corollary 1 (-calculus Completeness for -TR)**.** *The -calculus together with standard probability axioms is complete for the task of -TR.*

Proof. For the -operator, suppose $Q^k[\mathbf{A}_i;\ _{\mathbf{z}}] = P^k(\mathbf{a}_i;\ _{\mathbf{z}(\mathbf{A})})$ where $\mathbf{A}_i$ contains no element

in $\mathbf{Z} \subseteq \mathbf{X}$ is given, then

$$P^k(\mathbf{a}_i;\ \mathbf{z}_{(\mathbf{A})}) = P^k(\mathbf{a}_i;\ \mathbf{z}_{\setminus \mathbf{A}_i};\ \mathbf{v}_{\cap \mathbf{A}_i} = (\mathbf{v} \cap \mathbf{a}_i)) \tag{D.39}$$

$$= P^k(\mathbf{a}_i;\ \mathbf{v}_{\cap \mathbf{A}_i} = (\mathbf{v} \cap \mathbf{a}_i)) \tag{D.40}$$

$$= P^k(\mathbf{a}_i;\ \mathbf{x}_{\setminus \mathbf{A}_i};\ \mathbf{v}_{\cap \mathbf{A}_i} = (\mathbf{v} \cap \mathbf{a}_i)) \tag{D.41}$$

$$= P^k(\mathbf{a}_i;\ \mathbf{x}_{(\mathbf{A})}). \tag{D.42}$$

For the -operator, suppose no variable in $\mathbf{A}_i$ is in ^{k}, then in $G_{\overline{\mathbf{V} \cap \mathbf{A}_i}}$ any path between an T a variable in $\mathbf{A}_i$ is cut. Then

$$P^k(\mathbf{a}_i;\ \mathbf{x}_{(\mathbf{A})}) = P\ (\mathbf{a}_i;\ \mathbf{x}_{(\mathbf{A})}). \tag{D.43}$$

Having shown that all of the , , and operations follow from -calculus and standard probability axioms, it follows that any execution of C-INFER can be written as a -calculus derivation. Consequently, the completeness of C-INFER for -*TR* (theorem 6) implies the completeness of -calculus for the same task. □

Appendix E: Details on Examples of Soft-Transportability

E.1 Details on Example 19

First, recall the parametrization for the counterexample. Let

$$M^i = \langle \mathbf{V}; \mathbf{U}; F^i = \{f_{X_1}; f_Z^i; f_{X_2}; f_Y\}; P(\mathbf{U})\rangle; \tag{E.1}$$

where $\mathbf{V} = \{X_1; Z; X_2; Y\}$, $\mathbf{U} = \{U_{X_1Z}; U_Z; U_{X_2}; U_Y\}$ all $\mathbf{U}$ are binary, $U_{X_1Z}; U_Z$ and U_{X_2} are fair coins, and $P(U_Y = 1) = 3/4$, and

$$f_{X_1} : X_1 \leftarrow U_{X_1Z};$$

$$f_Z^0 : Z \leftarrow \begin{cases} X_1 \oplus U_Z & \text{if } X_1 = U_{X_1Z} \\ 0 & \text{otherwise} \end{cases};$$

$$f_Z^1 : Z \leftarrow \begin{cases} X_1 \oplus U_Z & \text{if } X_1 = U_{X_1Z} \\ 1 & \text{otherwise} \end{cases};$$

$$f_{X_2} : X_2 \leftarrow X_1 \oplus Z \oplus U_{X_2};$$

$$f_Y : Y \leftarrow X_1 \oplus X_2 \oplus U_Y;$$

and $\oplus$ is the binary *xor* operator.

The chosen conditional function for intervening X_2 is $g(X_1; Z) = X_1 \oplus Z$. The models $M^0{}_{x_1 x_2}$ and $M^1{}_{x_1 x_2}$. are identical to M^0 and M^1 except for the replacement f_{x_1} and f_{x_2}

as

$$f_{x_1} : X_1 \quad x_1;$$

$$f_{x_2} : X_2 \quad X_1 \quad Z:$$

Without any intervention, X_1 is always equal to $U_{x_1 z}$ so f_z^0 and f_z^1 coincide under observation and the models induce the same $P(\mathbf{V})$.

For simplicity denote $P(:;\ _{X_1};\ _{X_2}) = P\ (:)$ so that the target effect is $P^i(y)$. Also, to avoid confusion with summation indices, let x_1^0 be the constant fixed by $\ _{X_1} = do(X_1 = x_1^0)$. Since all models share the same set of $\mathbf{U}$ with the same distribution $P(\mathbf{U})$, eq. (2.16)

$$P^i(y) = \sum_{x_1;z;x_2} P^i(\mathbf{v}) \tag{E.2}$$

$$= \sum_{x_1;z;x_2;\mathbf{u}} P^i(x_1)P^i(z \mid x_1; u_{x_1 z}; u_z)P^i(x_2 \mid x_1; z)P^i(y \mid x_1; x_2; u_y)P(\mathbf{u}): \tag{E.3}$$

In both $M^i{}_{X_1\ X_2}$, $i = 0; 1$, $X_1 = x_1^0$ and $X_2 = X_1 \quad Z$, hence any event not consistent with these assignments has probability zero, and the corresponding terms can be removed from the summation

$$P^i(y) = \sum_{z;\mathbf{u}} P^i(z \mid x_1^0; u_{x_1 z}; u_z)P^i(y \mid x_1^0; X_2 = x_1^0 \quad z; u_y)P(\mathbf{u}): \tag{E.4}$$

The probability $P^i(y \mid x_1^0; X_2 = x_1^0 \quad z; u_y)$ is 1 if $y = x_1^0 \quad x_2 \quad u_y$, equivalently $y = z \quad u_y$. Remove from the sum terms where $U_y \neq Z \quad Y$:

$$P^i(y) = \sum_{z;u_{x_1 z};u_z} P^i(z \mid x_1^0; u_{x_1 z}; u_z)P(u_{x_1 z}; u_z)P(U_y = y \quad z): \tag{E.5}$$

The probability $P^i(z \mid x_1^0; u_{x_1 z}; u_z)$ is 1 if $(x_1^0 = u_{x_1 z}) \wedge (z = x_1^0 \quad u_z)$, or $(x_1^0 = u_{x_1 z}) \wedge (z =$

i); and it is zero otherwise. Accordingly

$$P^i(y) = \sum_z (P(U_{x_1 z} = x_1^0)P(U_z = x_1^0 \quad z) + 1[z = i]P(U_{x_1 z} \neq x_1^0))P(U_y = y \quad z) \tag{E.6}$$

$$= \sum_z (\tfrac{1}{4} + 1[z = i]\tfrac{1}{2})P(U_y = y \quad z) \tag{E.7}$$

$$= \sum_z \tfrac{1}{4}P(U_y = y \quad z) + \sum_z 1[z = i]\tfrac{1}{2}P(U_y = y \quad z) \tag{E.8}$$

$$= \tfrac{1}{4} + \sum_z 1[z = i]\tfrac{1}{2}P(U_y = y \quad z) \tag{E.9}$$

$$= \tfrac{1}{4} + \tfrac{1}{2}P(U_y = y \quad i): \tag{E.10}$$

Consequently $P^1(Y = 1) = 5/8$ while $P^2(Y = 1) = 3/8$.

E.2 Detailed Derivation of a -TR Instance

Here we provide a detail derivation of the causal effect $P(y; \ _X)$ for example 22 in section 5.4. All graphs we refer to are shown in fig. E.1.

First by marginalization:

$$P(y; \ _X) = \sum_{r;x;z} P(yjz; x; r; \ _X)P(zjx; r; \ _X)P(xjr; \ _X)P(r; \ _X): \tag{E.11}$$

We will transport factor by factor starting from the last.

By rule 3 and the separation $(R \perp\!\!\!\perp X)$ in $G_{\ _X\overline{X}}$ and $G_{\overline{X}}$, we have

$$P(r; \ _X) = P(r); \tag{E.12}$$

which is estimable from the input distribution $P(\mathbf{V})$.

The factor $P(xjr; \ _X)$ is determined by $\ _X$ (and the policy's specification).

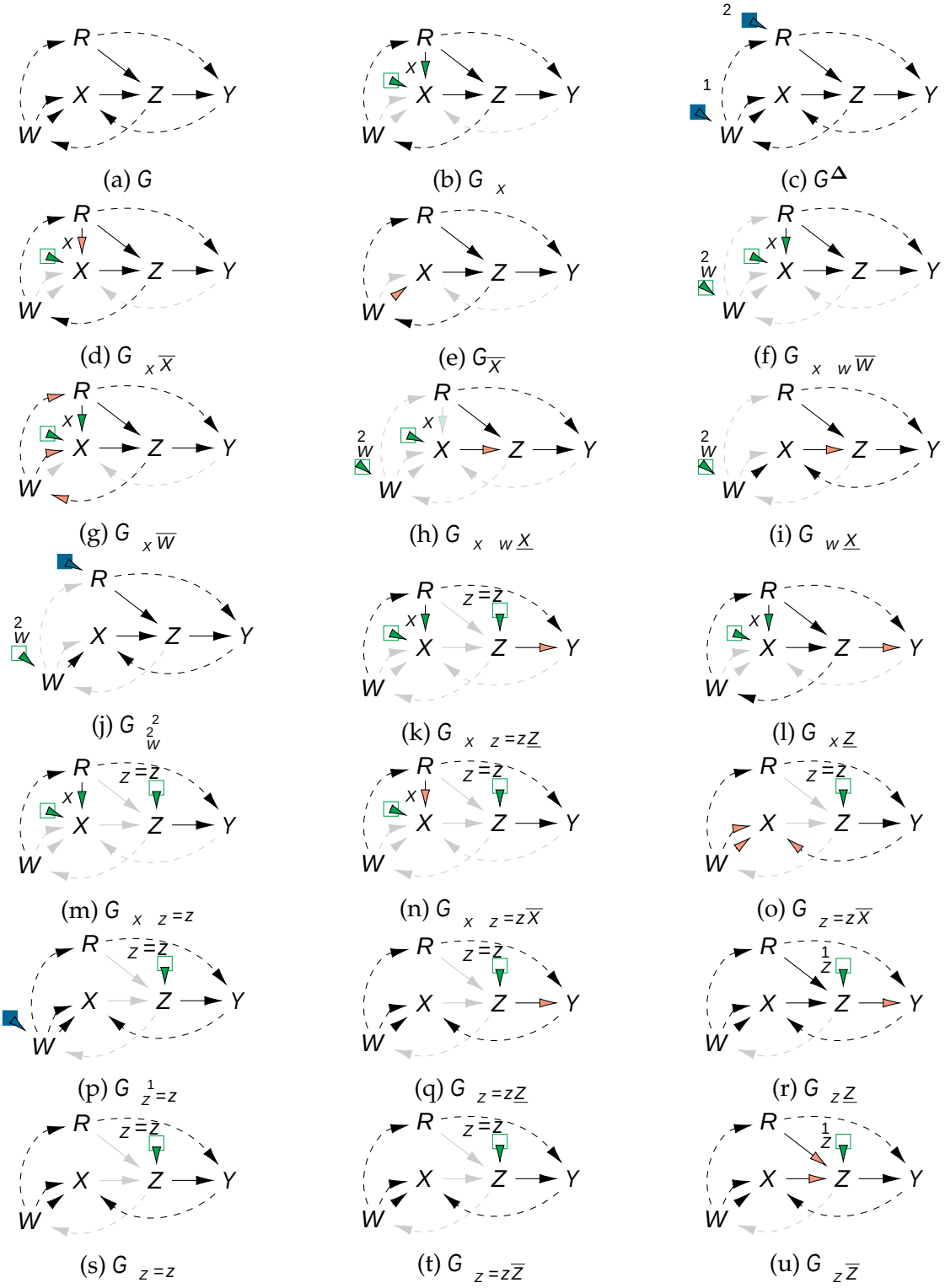


Figure E.1: Graphs used in the derivation for example 22. Edges added by intervention are shown in green, those removed by intervention in grey, and those cut due to the application of the rule are shown in faded red.

For the second factor,

$$P\ (zjx; r;\ \ _X) = P\ (zjx; r;\ \ _X;\ \ _W) \tag{E.13}$$

by rule 3 and $(Z \perp\!\!\!\perp W \mid X; R)$ in $G_{\ _X\ \ _W\overline{W}}$ and $G_{\ _X\overline{W}}$. Then, by rule 2 and $(Z \perp\!\!\!\perp X \mid R)$ in $G_{\ _X\ \ _W\underline{X}}$ and $G_{\ _W\underline{X}}$, we obtain

$$P\ (zjx; r;\ \ _X) = P\ (zjx; r;\ \ _W): \tag{E.14}$$

From the graph $G^{\ 2}_{\ _W}$, $(Z \perp\!\!\!\perp T_r \mid R; X)$ hence

$$P\ (zjx; r;\ \ _W) = P^2(zjx; r;\ \ _W) \tag{E.15}$$

estimable from the given $P^2(\mathbf{V};\ \ _W = \ ^2_W)$.

For the first factor is equal to

$$P\ (yjz; x; r;\ \ _X;\ \ _Z = z) \tag{E.16}$$

by rule 2 and $(Y \perp\!\!\!\perp Z \mid X; R)$ in $G_{\ _X\ \ _{Z=z}\underline{Z}}$ and $G_{\ _X\underline{Z}}$. We remove the observed X by rule 1 and $(Y \perp\!\!\!\perp X \mid Z; R)$ in $G_{\ _X\ \ _{Z=z}}$.

$$P\ (yjz; r;\ \ _X;\ \ _Z = z): \tag{E.17}$$

Then, by rule 3 and $(Y \perp\!\!\!\perp X \mid Z; R)$ in $G_{\ _X\ \ _{Z=z}\overline{X}}$ and $G_{\ _{Z=z}\overline{X}}$ this is equal to

$$P\ (yjz; r;\ \ _Z = z): \tag{E.18}$$

At this point we can transport this factor from ^{1} due to $(Y \perp\!\!\!\perp T_w \mid Z; R)$ in $G_{\frac{1}{Z}=z}$:

$$P^1(yjz; r; \ _Z): \tag{E.19}$$

We sum over X

$$\sum_{x'} P^1(yjz; x'; r; \ _Z=z)P^1(x' \mid z; r; \ _Z=z) \tag{E.20}$$

and use rule 2 with $(Y \perp\!\!\!\perp Z \mid X; R)$ in $G_{\ _{Z=z\underline{Z}}}$ and $G_{\ _{z\underline{Z}}}$ to exchange $\ _Z=z$ with $\ _Z=\ _{\frac{1}{Z}}$

$$\sum_{x'} P^1(yjz; x'; r; \ _Z)P^1(x' \mid z; r; \ _Z=z): \tag{E.21}$$

Due to $(X \perp\!\!\!\perp Z)$ in $G_{\ _{Z=z}}$ the observation of Z can be removed in the second factor in the sum:

$$\sum_{x'} P^1(yjz; x'; r; \ _Z)P^1(x' \mid r; \ _Z=z): \tag{E.22}$$

Note that $(X \perp\!\!\!\perp Z)$ in $G_{\ _{Z=z\overline{Z}}}$ and $G_{\ _{z\overline{Z}}}$, therefore by rule 3 we get

$$\sum_{x'} P^1(yjz; x'; r; \ _Z)P^1(x' \mid r; \ _Z); \tag{E.23}$$

resulting in a sum estimable from $P^1(\mathbf{V}; \ _Z)$.

Putting all together, we have:

$$P \ (y; \ _X) = \sum_{r;x;z} \underbrace{\sum_{x'} P^1(yjz; x'; r; \ _Z)P^1(x'jr; \ _Z)}_{\text{from } \frac{1}{Z} \text{ in } {}^{1}} \underbrace{P^2(zjx; r; \ _W)}_{\text{from } \frac{2}{W} \text{ in } {}^{2}} \underbrace{P \ (xjr; \ _X)}_{\text{def. } {}_X} \underbrace{P \ (r)}_{\text{from}} : \tag{E.24}$$

Appendix F: Adjustment Criteria

F.1 Proof of Generalized Adjustment Criterion (Selection bias)

In order to prove the theorem, we will magnify the causal diagram as described in [89], that is, we will replace every bidirected arrow connecting variables A and B with an observable variable $C_{A;B}$ that points to the pair previously connected by the bidirected arrow. Also, we will introduce a new mediator in every arrow leaving from any variable in $\mathbf{Y}$, that is, every edge of the form $Y \dashrightarrow A$ is replaced with $Y \dashrightarrow C_A \dashrightarrow A$, where $Y \in \mathbf{Y}$ and $A \in \mathbf{V}$. Let $\mathbf{C}$ be the set of all new variables introduced by the magnification process and $G[\mathbf{C}]$

Let any set with the subscript $\mathbf{nd}$ denote all the variables in such set that are not descendants of any variable in $\mathbf{X}$, that is, for any set $\mathbf{A}$ let $\mathbf{A}_{\mathbf{nd}} = \{A \in \mathbf{A} \mid A \notin De(\mathbf{X})\}$. Analogously, the subscript $\mathbf{d}$ will denote all the variables in that are descendants of any variable in $\mathbf{X}$.

We show that the causal effect can be derived from the available data. In order to perform the steps we will define several subsets of $\mathbf{Z}$ such as:

- $\mathbf{Z}^{\mathbf{S}} = \{Z \in \mathbf{Z}^{\mathbf{M}} \mid (Z \not\perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}})\}$.
- $\mathbf{L_1}$ to be all the variables in $(\mathbf{V} \cup \mathbf{C}) \cap (\mathbf{Z} \cup \mathbf{X} \cup \mathbf{Y})$ that:
  1. Are d-connected to $\mathbf{Y}$ given $\mathbf{Z}^{\mathbf{T}}; \mathbf{Z}^{\mathbf{S}}$ in $G \cap \mathbf{X}$
  2. Are not descendants of $\mathbf{X}$
  3. Are ancestors of some $Z \in \mathbf{Z}^{\mathbf{T}} \cup \mathbf{Z}^{\mathbf{S}}$
- $\mathbf{L_2}$ be defined as all variables in $(\mathbf{V} \cup \mathbf{C}) \cap (\mathbf{Z} \cup \mathbf{X} \cup \mathbf{Y})$ that:

1. Are d-connected to $\mathbf{Y}$ given $\mathbf{Z}$ in $G_{\overline{\mathbf{X}}}$
2. Are independent of $\mathbf{X}$ given $\mathbf{Z^S}$, $\mathbf{Z^T}$ and S on $G_{\overline{\mathbf{X}(\mathbf{Z^S};\mathbf{Z^T};S)}}$
3. Are ancestors of some variable in $\mathbf{Z}$.

- $\mathbf{Z^X} = \{Z \in \mathbf{Z} \mid (\mathbf{Z^T} \cup \mathbf{Z^S}) \not\perp (Z \in \mathbf{X} \mid \mathbf{Z^S}; \mathbf{Z^T}; S)_{G_{\overline{\mathbf{X}(\mathbf{Z^S};\mathbf{Z^T};S)}}}\}$
- $\mathbf{Z^Y} = \mathbf{Z} \setminus (\mathbf{Z^T} \cup \mathbf{Z^S} \cup \mathbf{Z^X})$.

lemma 22 (which is after the claims below) proves independences that will be used in the derivation. But before stating it, we will need the following claims:

Claim 3. *If there exist a path r_1 between some $X' \in \mathbf{X}$ and S where S is a descendant of X', such that r_1 does not contain any variable in $\mathbf{Z^T}$ and the conditions from definition 22 are satisfied, then r_1 exists in $G^{pbd}_{\mathbf{XY}}$.*

Proof. Path r_1 can be absent in $G^{pbd}_{\mathbf{XY}}$ only if the edge from X' towards S belongs to a proper causal path. For the sake of contradiction suppose this is the case, and let R be the variable at the other end of that edge. It follows that R is in a proper causal path that ends with some $Y'' \in \mathbf{Y}$, and does not contain any variable in $\mathbf{Z}$ for the satisfaction of condition (i). Since the path between R and S is a subpath of r_1 it does not contain any $\mathbf{Z^T}$ and if it contains any other variable in $X'' \in \mathbf{X}$ restart the argument with $X' = X''$. Then, the path formed between S and Y' passing through R exists in $G^{pbd}_{\mathbf{XY}}$ and is active given $\mathbf{Z^T}$, contradicting condition (iii). Since we reached a contradiction the edge must be in the mentioned graph and r_1 as well. □

Claim 4. *If there exists a path r_1 between a variable $W' \in \mathbf{Z}$ and some $Y' \in \mathbf{Y}$ that does not contain any variable in $\mathbf{X}$ and is active given $\mathbf{Z^T}; \mathbf{Z^S}$. And there exists also, a path r_2 directed from W' and S that does not contain any $\mathbf{Z^T}; \mathbf{Z^S}$. Then, condition (iii) is violated.*

Proof. If r_2 contains some $\mathbf{X}$ let X' be the closest to W', then $X' \in An(S)$ and by Claim 3, the path between X' and S exists in $G^{pbd}_{\mathbf{XY}}$. Then, r_2 is active in $G^{pbd}_{\mathbf{XY}}$ given $\mathbf{Z^T}; \mathbf{Z^S}$.

Let r be the path between S and Y' composed with edges from r_1 and r_2. Path r exists in $G^{pbd}_{\mathbf{XY}}$ and is active given $\mathbf{Z^T}$, then condition (iii) is not satisfied. □

Claim 5. *If there exists a non-causal path p between some $X' \in \mathbf{X}$ and some $Y' \in \mathbf{Y}$ which is active in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$. And, such path p does not contain any other variable in $\mathbf{X}$ asides from X'. Then the conditions in definition 22 are not satisfied.*

Proof.

- *Path p has to be active given $\mathbf{Z^T}; \mathbf{Z^S}$ (without S).* This is because if not observing S closes p, it implies that S is the descendant of a collider W' in p, such that W' is active given S but inactive otherwise. Let r_1 be the path between W' and Y' and let r_2 bet the path between W' and S, then by virtue of Claim 4 condition (iii) is violated, a contradiction.

- Since p is a non-causal path, condition (ii) requires it to be closed given $\mathbf{Z} \cup \{S\}$. Then, there must exists some $Z \in \mathbf{Z} \setminus (\mathbf{Z^T} \cup \mathbf{Z^S})$ that closes p (note that S cannot block any path and particularly p). We will show that such Z cannot exists under the criterion's conditions.

- Suppose there exists a Z that blocks p. Since $Z \notin \mathbf{Z^S}$, there exist a path q_1 between S and Z that is active given $\mathbf{Z^T}$. Let q be the path between S and Y' formed using edges from q_1 and p.

  - *The path q_1 does not contain any variable in $\mathbf{Z^S}$.* Since q_1 is active given $\mathbf{Z^T}$, any variable in it is not independent of S given $\mathbf{Z^T}$, hence none of them can be in $\mathbf{Z^S}$.

  - According to condition (iii) q has to be blocked given $\mathbf{Z^T}$ in $G^{pbd}_{\mathbf{XY}}$. Since p does not contain any $\mathbf{X}$ (other than X') and it is active given $\mathbf{Z^T}; \mathbf{Z^S}$ the path q must be closed in $G^{pbd}_{\mathbf{XY}}$ because (1) Z is a collider in it, (2) there exists a collider in p that belongs to $\mathbf{Z^S}$, or (3) there exists some X'' in q_1 for which one or two edges in q_1 are present in G but not in $G^{pbd}_{\mathbf{XY}}$:

(1) *If Z is a collider in q, then Z is active because it has a descendant in* $\mathbf{Z^T}$ in $G^{pbd}_{\mathbf{XY}}$. The portion of p that goes from X' to Z has an edge coming out of Z for it to block p and be a collider in q at the same time. Regarding that portion:

- *it does not contain any variable in* $\mathbf{Z^S}$: Suppose it does, then there is a path between S to that variable going through Z that contradicts the definition for $\mathbf{Z^S}$.
- if it has a collider in between, it must be active given $\mathbf{Z^T}$, and it is a descendant of Z, implying that Z is active as well.
- *if it is directed from Z to X', then there exists a $Z'' \in \mathbf{Z^T}$ which is a descendant of Z*. We know that the edge incoming to X' exists in $G_{\overline{X}(\mathbf{Z^T};\mathbf{Z^S};S)}$ (by assumption of p), then X' must have a descendant in $\mathbf{Z^T} \cup \mathbf{Z^S} \cup \{S\}$. Any descendant of X' in $\mathbf{Z^S}$ is d-connected to S given $\mathbf{Z^T}$ with a path passing through X' and Z unless some variable in $\mathbf{Z^T}$ is in between that descendant and X', which means that X' has to have a descendant $Z'' \in \mathbf{Z^T} \cup \{S\}$.

  If the descendant of X' is S without any $\mathbf{Z^T}$ in between, let the path between X' and S be called r_1 which exists $G^{pbd}_{\mathbf{XY}}$ by Claim 3. Let p' be the path from S to Y' formed joining r_1 and p. Path p' exists in $G^{pbd}_{\mathbf{XY}}$ and cannot contain any variable in $\mathbf{Z^S}$ because such variable would not satisfy the independence that defines $\mathbf{Z^S}$. Therefore, p' is active given $\mathbf{Z^T}$ alone and witnesses a contradiction to condition (iii). As a consequence $Z'' \notin \mathbf{Z^T}$.

  The edge outgoing from X' towards Z'' is not in a proper causal path. Assume for the sake of contradiction that it is, then let R be the variable at the other end of that edge (possibly Z'' itself). This means that R is in a proper causal path and has Z'' as descendant, contradicting condition (i).

As a result, $Z'' \in \mathbf{Z_T}$ is a descendant of Z in $G^{pbd}_{\mathbf{XY}}$, which make Z an active collider.

(2) *If Z is not an inactive collider in q (given $\mathbf{Z_T}$), then q does not contain any variable in $\mathbf{Z_S}$*. If this was the case, that variable would be d-connected to S given $\mathbf{Z_T}$ which contradicts the definition of $\mathbf{Z_S}$.

(3) *If neither Z nor some $\mathbf{Z_S}$ block q, there must exists a X'' (as defined before), but it is not possible under the criterion conditions*: For X'' to disconnect q in $G^{pbd}_{\mathbf{XY}}$, the path q_1 should have one of the following structures:

* $Z \to X'' \to S$ or $Z \to X'' \to \cdots \to S$ *where the outgoing edge from X'' belongs to a proper causal path*: Suppose the edge towards Z is in a proper causal path and let R be the variable at the other end of that edge (possibly Z itself). Then either $Z \in De(R)$ or there exists some variable in $\mathbf{Z_T} \cap De(R)$ in the path between X'' and Z, in both cases condition (i) is violated because there is a descendant of R that is observed while R is in a proper causal path.

  Now suppose the edge towards S is in a proper causal path and let R be the variable in the other end of that edge. Then R has to be an ancestor of S otherwise there is a collider in between (possibly R itself) that must be active given $\mathbf{Z_T}$ which implies that R is in $\mathbf{Z_T}$ or has a descendant on it, violating condition (i). Then X'' is an ancestor of S and the path between them has no $\mathbf{Z_T}$. By Claim 3 the path between them, hence the outgoing edge, exists in $G^{pbd}_{\mathbf{XY}}$.

* $Z \leftarrow \cdots \leftarrow X'' \to \cdots S$: In this case Z would be independent of S given $\mathbf{Z_T}$ because of X'' not being active, contradicting our assumption that $Z \notin \mathbf{Z_S}$ because of path q_1.

□

Lemma 22. *Suppose that in the causal diagram G there are sets of variables $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{Y}$, such that $\mathbf{Z}$ is admissible by the criterion in definition 22 relative to $\mathbf{X}$ and $\mathbf{Y}$. Then, the following independences hold in the magnification of G:*

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}_{\mathbf{d}}^{\mathbf{T}}, \mathbf{Z}_{\mathbf{d}}^{\mathbf{S}} \mid \mathbf{L}_1, \mathbf{Z}_{\mathbf{nd}}^{\mathbf{T}}, \mathbf{Z}_{\mathbf{nd}}^{\mathbf{S}}, \mathbf{X})_{G_{\overline{\mathbf{X}}}} \tag{F.1}$$

$$(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}}, \mathbf{X})_{G_{\overline{\mathbf{X}}}} \tag{F.2}$$

$$(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}}, \mathbf{Z}^{\mathbf{S}}, \mathbf{X})_{G_{\overline{\mathbf{X}}}} \tag{F.3}$$

$$(\mathbf{L}_1 \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}^{\mathbf{T}}, \mathbf{Z}^{\mathbf{S}})_{G_{\overline{\mathbf{X}(\mathbf{Z}^{\mathbf{T}}, \mathbf{Z}^{\mathbf{S}})}}} \tag{F.4}$$

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}^{\mathbf{Y}} \mid \mathbf{Z}^{\mathbf{T}}, \mathbf{Z}^{\mathbf{S}}, \mathbf{L}_2, \mathbf{Z}^{\mathbf{X}}, \mathbf{X}, S)_{G_{\overline{\mathbf{X}}}} \tag{F.5}$$

$$(\mathbf{L}_2 \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}, S)_{G_{\overline{\mathbf{X}(\mathbf{Z}, S)}}} \tag{F.6}$$

Proof. We will go over each independence and show that it holds:

1. $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}_{\mathbf{d}}^{\mathbf{T}}, \mathbf{Z}_{\mathbf{d}}^{\mathbf{S}} \mid \mathbf{L}_1, \mathbf{Z}_{\mathbf{nd}}^{\mathbf{T}}, \mathbf{Z}_{\mathbf{nd}}^{\mathbf{S}}, \mathbf{X})_{G_{\overline{\mathbf{X}}}}$: Suppose it does not hold, then there exists a path p between $Z^0 \in \mathbf{Z}_{\mathbf{d}}^{\mathbf{S}} \cup \mathbf{Z}_{\mathbf{d}}^{\mathbf{T}}$ and $Y^0 \in \mathbf{Y}$. The path p cannot contain any variable in $\mathbf{X}$ because in the graph $G_{\overline{\mathbf{X}}}$ when $\mathbf{X}$ is observed, any path containing it is closed. The path p should have one of the following structures:

   (a) $Z^0 \rightarrow \cdots \rightarrow Y^0$: Since Z^0 is a descendant of $\mathbf{X}$ by definition, then it belongs to a proper causal path and contradicts cond. (i)

   (b) $Z^0 \leftarrow \cdots \rightarrow Y^0$: Here Z^0 is a descendant of Y^0, so Y^0 is not a descendant of $\mathbf{X}$ otherwise Z^0 contradicts cond. (i) because it is a descendant of Y^0 which is part of a proper causal path. But if Y^0 is not a descendant of $\mathbf{X}$, the child of Y^0 in p is a node added in the magnification process and satisfies the definition of $\mathbf{L}_1$ (it is always d-connected to Y^0, it is not a descendant of $\mathbf{X}$ and it is an ancestor of Z^0) therefore p is blocked.

(c) $Z^0 \; ! \; W \; Y^0$: There is a collider W in p that belongs or has a descendant $W^0 \; 2 \; \mathbf{L_1} \; [\; \mathbf{Z_{nd}^T} \; [\; \mathbf{Z_{nd}^S}$ where W could be equal to W^0, such that p is active. Without loss of generality assume that W is the closest of such colliders to Y^0 in p. Let L^0 be the parent of W in the section of p that goes from W to Y^0, we want to show that L^0 belongs to $\mathbf{L_1}$ and blocks p.

First, note that L^0 cannot be Y^0 itself because the edge between Y^0 and W would be replaced with a mediator during the magnification process, making the mediator the parent of W in p. Second, we can assure that the path between W^0 and Y^0 does not contain any variable $Z \; 2 \; \mathbf{Z_d^T} \; [\; \mathbf{Z_d^S}$ because if it does, Z has to be an ancestor of Y^0 or W^0, in the first case violating condition (i) and in the second W^0 would also be a descendant of $\mathbf{X}$ which is not possible given its definition. Third, L^0 is d-connected to Y^0 given $\mathbf{Z^T}; \mathbf{Z^S}$ in $G \; n \; \mathbf{X}$ because the path from W^0 to Y^0 is active given $\mathbf{Z_{nd}^T} \; [\; \mathbf{Z_{nd}^S}$ and does not contain any $\mathbf{Z_d^T} \; [\; \mathbf{Z_d^S}$ and does not contain any $\mathbf{X}$. L^0 is also an ancestor of all the descendant of W including W^0, if W^0 belongs to $\mathbf{L_1}$, it has, by definition, a descendant in $\mathbf{Z^T} \; [\; \mathbf{Z^S}$ that are also descendants of L^0. Finally L^0 is not a descendant of $\mathbf{X}$ otherwise W^0 is also descendant of $\mathbf{X}$ which is not possible by its definition. Therefore $W^0 \; 2 \; \mathbf{L_1}$.

(d) $Z^0 \; ! \; Y^0$: Here the path is not completely directed in any direction but does not contain any collider. Let L^0 be the common ancestor of Z^0 and Y^0 in p. Note that L^0 satisfy the definition of $\mathbf{L_1}$ and closes p as follows: L^0 is always an ancestor of Z^0, neither L^0 nor any $Z \; 2 \; \mathbf{Z}$ in the path between L^0 and Y^0 is a descendant of $\mathbf{X}$, otherwise condition (i) is violated because they would lie in a proper causal path with descendants in $\mathbf{Z}$. By assumption L^0 is d-connected to Y^0 given $\mathbf{Z_{nd}^T}; \mathbf{Z_{nd}^S}$, and since no descendant of $\mathbf{X}$ (i.e. $\mathbf{Z_d^T}; \mathbf{Z_d^S}$) is on the path between them, L^0 and Y^0 are connected given $\mathbf{Z^T}; \mathbf{Z^S}$ in $G n \mathbf{X}$. Therefore, $L^0 \; 2 \; \mathbf{L_1}$ and p is closed.

2. $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{X})_{G_{\overline{\mathbf{X}}}}$: From condition (iii) we have:

$$(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T})_{G^{pbd}_{\mathbf{XY}}} \tag{F.7}$$

$$=) (\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T})_{G^{pbd}_{\mathbf{XY}\overline{\mathbf{X}}}} \tag{F.8}$$

$$=) (\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{X})_{G^{pbd}_{\mathbf{XY}\overline{\mathbf{X}}}} \tag{F.9}$$

$$=) (\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{X})_{G_{\overline{\mathbf{X}}}} \tag{F.10}$$

eq. (F.8) follows because removing incoming edges to $\mathbf{X}$ never introduces dependencies. Provided that no variable in $\mathbf{X}$ has incoming edges, introducing $\mathbf{X}$ to the set of observed variables may never compromise a previously established independence hence (F.9) follows. Finally, comparing the graphs $G_{\overline{\mathbf{X}}}$ and $G^{pbd}_{\mathbf{XY}\overline{\mathbf{X}}}$ we can see that the former could possibly have edges that are not in the second. Those edges are those that have tails in $\mathbf{X}$ and do not belong to a proper causal path. Since $\mathbf{X}$ is being observed in the independence any new path including those edges is always block, therefore independence (F.10) is implied.

3. $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{Z^S}; \mathbf{X})_{G_{\overline{\mathbf{X}}}}$: From the previous independence and the definition of $\mathbf{Z^S}$ we have:

$$(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{X})_{G_{\overline{\mathbf{X}}}} \wedge (\mathbf{Z^S} \perp\!\!\!\perp S \mid \mathbf{Z^T}) \tag{F.11}$$

$$=) (\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{X})_{G_{\overline{\mathbf{X}}}} \wedge (\mathbf{Z^S} \perp\!\!\!\perp S \mid \mathbf{X}; \mathbf{Z^T})_{G_{\overline{\mathbf{X}}}} \tag{F.12}$$

$$=) (\mathbf{Y}; \mathbf{Z^S} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{X})_{G_{\overline{\mathbf{X}}}} \tag{F.13}$$

$$=) (\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^S}; \mathbf{Z^T}; \mathbf{X})_{G_{\overline{\mathbf{X}}}} \tag{F.14}$$

Here statement (F.12) follows from the fact that conditioning on $\mathbf{X}$ while transforming the graph from G to $G_{\overline{\mathbf{X}}}$ can only add more independences, but does not remove any of the existent ones. Statement (F.13) follows from the composition axiom that holds whenever d-separation holds. Finally, statement (F.14) follows by weak union.

4. $(\mathbf{L_1} \not\perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z^T}; \mathbf{Z^S})_{G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S})}}}$: Assume for the sake of contradiction that this does not hold. Then, fix a path p_1 from some $L' \in \mathbf{L_1}$ to some $X' \in \mathbf{X}$ given $\mathbf{Z^T}; \mathbf{Z^S}$ in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S})}}$ ($X' \quad L'$). Without loss of generality assume that X' is the closest variable in $\mathbf{X}$ to L' in p_1, if it is not, restart the argument with that other $\mathbf{X}$ instead. By definition, L' is d-connected to some $Y' \in \mathbf{Y}$ by some path p_2 in $G \setminus \mathbf{X}$ given $\mathbf{Z^T}; \mathbf{Z^S}$ ($L' \quad Y'$). Let p be the path between X' and Y' formed using edges in p_1 and p_2 in G, also let W' be the closest node to L' that belongs to p, possibly L' itself.

   - *If W' is connected to L' by a path that starts with an edge going out from W' (i.e. $W' \rightarrow \quad L'$), then W' has a descendant $Z' \in \mathbf{Z^T} \cup \mathbf{Z^S}$*: Either W' is ancestor of L' and, by extension, of some $Z' \in \mathbf{Z^T} \cup \mathbf{Z^S}$ by definition of $\mathbf{L_1}$ (i.e. $W' \rightarrow \quad \rightarrow L' \rightarrow \quad \rightarrow Z'$) or there is an active collider connecting W' and L' in the very same set, such collider is a descendant of W' ($W' \rightarrow \quad \rightarrow Z' \quad L'$).
   - *p does not contain any variable in $\mathbf{X}$ except at the endpoint*: This is because p_1 does not include any other $\mathbf{X}$, and p_2 is active in $G \setminus \mathbf{X}$.
   - *p is not a proper causal path.* Suppose it is a proper causal path, and consider the relationship between L' and W'.
     - If $W' = L'$ then $L' \in De(\mathbf{X})$ contradicting the definition of $\mathbf{L_1}$.
     - If the path between W' and L' has an edge pointing into W' then W' is a collider in p_1 implying that X' and L' are disconnected unless W' has a descendant in $\mathbf{Z^S} \cup \mathbf{Z^T}$. Since W' is assumed to be in a proper causal path, such descendant violates condition (i).
     - If the path between L' and W' has edges outgoing from W' then Z' is a descendant of W', and since W' is in a proper causal path, Z' contradicts condition (i).
   - *W' is not an inactive collider in p (given $\mathbf{Z^T}; \mathbf{Z^S}$)*: If W' is a collider, it is connected to L' with an incoming or outgoing edge. If the edge is incoming to W', then W'

has to be active for p_1 to be active. If the edge is going out from W' or $W' = L'$, it follows that W' is active in p because it is an ancestor of Z'.

- *By Claim 5 the criterion is not satisfied.* Note that p is a non-causal path and it is active in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S})}}$ given $\mathbf{Z^T}; \mathbf{Z^S}$ and does not contain any variable in $\mathbf{X}$ except for X'. The path p also exists in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ because it contains the same or more edges than $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S})}}$. And p is active given $\mathbf{Z^T}; \mathbf{Z^S}; S$ also because observing S cannot close it. Therefore, Claim 5 applies to p.

5. $(\mathbf{Y} \not\perp\!\!\!\perp \mathbf{Z^Y} \mid \mathbf{Z^T}; \mathbf{Z^S}; \mathbf{L_2}; \mathbf{Z^X}; \mathbf{X}; S)_{G_{\overline{\mathbf{X}}}}$: Assume the independence does not hold and fix a path p_1 that connects some $Y' \in \mathbf{Y}$ to $Z' \in \mathbf{Z^Y}$ given $\mathbf{Z^T}; \mathbf{Z^S}; \mathbf{L_2}; \mathbf{Z^X}; \mathbf{X}; S$ in $G_{\overline{\mathbf{X}}}$. Without loss of generality assume that Z' is the closest of such variables to Y' in p_1.

   - *p_1 is active given $\mathbf{Z^T}; \mathbf{Z^S}; \mathbf{L_2}; \mathbf{Z^X}; S$ (without $\mathbf{X}$) in $G_{\overline{\mathbf{X}}}$.* Path p_1 does not contain any variable in $\mathbf{X}$ except at the endpoint because p_1 is active in $G_{\overline{X}}$ given $\mathbf{X}$, which means that it would be blocked or missing an edge if it contains $\mathbf{X}$.

   - *p_1 is active given $\mathbf{Z^T}; \mathbf{Z^S}; \mathbf{Z^X}; S$ (without $\mathbf{L_2}$) in $G_{\overline{\mathbf{X}}}$.* Path p_1 does not contain any variable in $\mathbf{L_2}$ or any collider activated exclusively by $\mathbf{L_2}$ and not by $\mathbf{Z^T}$, $\mathbf{Z^S}$, $\mathbf{Z^X}$, S.

     Suppose for the sake of contradiction that there is any variable $L' \in \mathbf{L_2}$ activating p_1. Since the path is assumed to be active given $\mathbf{L_2}$, L' may only be a collider in p_1 or it is the descendant of some Q which is a collider in p_1 that is active given L' but not $\mathbf{Z^T}; \mathbf{Z^S}; \mathbf{Z^X}; S$. It must be the case that the path between Q and L' does not contain any variable in $\mathbf{Z^T}; \mathbf{Z^S}; \mathbf{Z^X}; S$. Furthermore, it cannot contain any variable in $\mathbf{Z^Y}$ either because it would not be independent of $\mathbf{X}$ in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$ because its ancestor in $\mathbf{Z^Y}$ is not.

     Let Q' denote either L' if it is in p_1 or Q in the second case. Let R be the parent of Q' in the portion of p_1 that goes towards Y'. Note that R cannot be a collider in this path and $R \neq Y'$ because if $Y' \to Q'$ was an edge in G, a new mediator

was introduced during magnification. Furthermore, $R \in \mathbf{L_2}$: if L' satisfies the first part of the definition of $\mathbf{L_2}$, R which is an ancestor of L' also satisfies it. Even if L' is d-connected to a variable in $\mathbf{Y}$ other than Y', and behaves as a collider between R and that variable, L' is active given $\mathbf{Z}$. For the second part, R has to be independent of $\mathbf{X}$ as stated in the definition of $\mathbf{L_2}$, otherwise L' would not satisfy this either. For the third part, $L' \in An(\mathbf{Z})$ and $R \in An(L')$ then $R \in An(\mathbf{Z})$. As a consequence of $R \in \mathbf{L_2}$, p_1 is blocked by it, which is a contradiction to our assumption, and the conclusion follows.

- *p_1 is active given $\mathbf{Z^T}; \mathbf{Z^S}; S$ (without $\mathbf{Z^X}$) in $G_{\overline{\mathbf{X}}}$.* Follows from the fact that p_1 does not contain any variable in $\mathbf{Z^X}$. Suppose this is not true, then let $Z \in \mathbf{Z^X}$ be the closest of such variables to Y' in p_1. Z has to be an active collider for p_1 to be active. Let R be the parent of Z in the portion of p_1 that goes towards Y'. Note that $R \notin Y'$ because if $Y' \rightarrow Z$ was an edge in G, a new mediator was introduced during magnification and R would be that mediator. Then, $R \in \mathbf{L_2}$ and blocks p_1, because: first, R is d-connected to Y' through p_1 given $\mathbf{Z}$ unless S is a collider in between, but then independence (F.3) is violated (no variables in $\mathbf{Z^X} \cup \mathbf{Z^Y}$ are in this portion because we assumed Z and Z' were the closest to Y' in this path). Second R is independent of $\mathbf{X}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$ on $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ else Z would not satisfy this independence either, which is not the case by definition. Third, R is the ancestor of Z. Since R would block p_1, Z cannot exists in p_1.

- *p_1 is active given $\mathbf{Z^T}; \mathbf{Z^S}$ (without S) in $G_{\overline{\mathbf{X}}}$.* This is because p_1 does not contain S. Suppose it does, then the subpath between S and Y' violates independence (F.3).

- Since Z' does not belong to $\mathbf{Z^X}$, there exists a path p_2 that connects Z' to some $X' \in \mathbf{X}$ in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$. Assume, without loss of generality, that X' is the closest variable in $\mathbf{X}$ to Z' in the path p_2. Let p the path between X' and Y' that uses edges in p_1 and p_2 and let W' be the closest node to Z' in p, possibly Z' itself.

- *p is not a causal path.* For the sake of contradiction suppose p is a causal path, since p_1 exists in $G_{\overline{\mathbf{X}}}$ and p_2 only contains X' from $\mathbf{X}$, p is also a proper causal path. Now, consider the path between W' and Z':
  - if $W' = Z'$ we have a contradiction to condition (i).
  - if it starts with an incoming edge, then W' is an active collider in p_2, with a descendant in $\mathbf{Z^T} \cup \mathbf{Z^S} \cup \{S\}$. If the descendant is specifically in $\mathbf{Z^T} \cup \mathbf{Z^S} \setminus \mathbf{Z}$ there is a violation to condition (i).

    If W' is an ancestor of S, let r_1 be the path between W' and Y', which cannot contain any $\mathbf{Z} \cup \mathbf{X}$ by condition (i) and definition of proper causal path. Also let r_2 be the path between W' and S which does not contain any $\mathbf{Z^T}; \mathbf{Z^S}$. Then, by Claim 4 condition (iii) is violated.
  - if it starts with an outgoing edge, then W' is an ancestor of Z' or a collider that is active in p_2 (i.e. $\mathbf{Z^T}; \mathbf{Z^S}; S$). If it is ancestor of Z' condition (i) is not satisfied. If ancestor of S the same argument as before applies again.
- *p_2 does not contain any variable in $\mathbf{Z^X} \cup \mathbf{L_2}$*: Both sets require the independence $(\mathbf{Z^X}; \mathbf{L_2} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z^T}; \mathbf{Z^S}; S)_{G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}}$. Any variable in p_2 does not satisfy that independence by definition of p_2.
- *If W' does not block p given $\mathbf{Z^T}; \mathbf{Z^S}; S$, then by Claim 5 the conditions of the criterion in definition 22 are violated.* To see this observe that if W' does not block p then it is active in the graph $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$ and does not contain $\mathbf{X}$ except for X', hence Claim 5 applies to it.
- *If W' blocks p given $\mathbf{Z^T}; \mathbf{Z^S}; S$, then the criterion in definition 22 is violated*: In this case W' has to be an inactive collider in p. If the edge that has Y' as endpoint in p is outgoing from Y', let Q be the variable introduced as a mediator during magnification. If the edge is incoming to Y', let Q be farthest ancestor of Y' in p. Note that Q cannot be W' itself because, even if Y' was the parent of W' in G, Q

is a mediator. Here Q is d-connected to Y' given $\mathbf{Z}$ (no variable in $\mathbf{Z^X}; \mathbf{Z^Y}$ is in p at all and any from $\mathbf{Z^T} \setminus \mathbf{Z^S}$ would block p_1). Also Q is an ancestor of some $\mathbf{Z}$ (because the portion of p from Q to W' is either directed and W' is ancestor of Z', or the subpath has a collider in $\mathbf{Z^T} \setminus \mathbf{Z^S}$). No incoming edge to Q is possible in this section because its neighbor would be the farthest ancestor of Y' instead of Q and if Q is the mediator the edge must be outgoing. Then Q will be in $\mathbf{L_2}$ unless it is not independent of $\mathbf{X}$ in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$. If $Q \in \mathbf{L_2}$, p_1 is closed by Q. Hence, Q must not satisfy this independence. Yet the reason is not because of the path p, where Q is independent of X' in this sense. It follows that there exists a path p_1'' between some $X'' \in \mathbf{X}$ and Q in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$, active given $\mathbf{Z^T}; \mathbf{Z^S}; S$. Without loss of generality assume X'' is the closest of such variables to Q in p_1''. Meanwhile, the path p_2'' from Q to Y' is open in $G_{\overline{\mathbf{X}}}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$. Let p' be the path formed between X'' and Y' by joining edges from p_1'' and p_2''. The path p' has one of the following structures:

- $X'' \quad Q \quad Y'$ (if Y' is a parent of Q then it is a mediator because of the magnification).
- $X'' \quad Q \; ! \quad ! \; Y'$
- $Q \quad Y' \quad X''$ (here Y' lies in p_1'')
- $Q \; ! \quad ! \; Y' \quad X''$ (here Y' lies in p_1'')

In the four cases the path p' is active in $G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}$ given $\mathbf{Z^T}; \mathbf{Z^S}; S$ and does not contain any $\mathbf{X}$ asides from X''. Then, Claim 5 provides that the existence of p' contradicts our assumption that criterion in definition 22 was satisfied, a contradiction.

6. $(\mathbf{L_2} \not\perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}; S)_{G_{\overline{\mathbf{X}(\mathbf{Z};S)}}}$: Suppose this does not hold. Then, there exists a path p_1 from $X' \in \mathbf{X}$ to $L' \in \mathbf{L_2}$ active given $\mathbf{Z}; S$ in $G_{\overline{\mathbf{X}(\mathbf{Z};S)}}$. Assume without loss of generality that X' is the variable in $\mathbf{X}$ closest to L' in the path p_1. By definition of $\mathbf{L_2}$ there is

also a path p_2 from L' to some $Y' \in \mathbf{Y}$ that is active when $\mathbf{Z}$ is observed in $G_{\overline{\mathbf{X}}}$. Let p be the path between X' and Y' formed using edges from p_1 and p_2 and let W' be the closest variable to L' that lies in p, possibly L' itself.

- *The only variable in* $\mathbf{X}$ *that p contains is* X'. This follows by the assumption that X' is the closest to L' in p_1 and the fact that p_2 cannot be active in $G_{\overline{\mathbf{X}}}$ if it contains any variable in $\mathbf{X}$.
- L' *has a descendant* $Z' \in \mathbf{Z}$ *by definition of* $\mathbf{L_2}$
- *If* W' *is connected to* L' *by a path that starts with an edge going out from* W' *(i.e.* $W' \rightarrow \cdots \rightarrow L'$*), then* W' *has a descendant* $Z \in \mathbf{Z} \cup \{S\}$: Either W' is ancestor of L' and, by extension, of Z' (i.e. $W' \rightarrow \cdots \rightarrow L' \rightarrow \cdots \rightarrow Z' = Z$) or there is an active collider in $\mathbf{Z} \cup \{S\}$ connecting W' and L', such collider is descendant of W' ($W' \rightarrow \cdots \rightarrow Z' \leftarrow \cdots L'$).
- *the path p is not causal*. Suppose it is causal, then by the previous argument it has to be a proper one too. Consider the relationship between L' and W'.
  - If $W' = L'$. By definition of $\mathbf{L_2}$, L' has a descendant in $\mathbf{Z}$ which contradicts condition (i) because W' is assumed to be in a proper causal path.
  - If the path between W' and L' has an edge pointing into W', then W' is a collider in p_1 implying that X' and L' are disconnected unless W' has a descendant in $\mathbf{Z} \cup \{S\}$. If the descendant is in $\mathbf{Z}$ it violates condition (i).
    If W' is an ancestor of S, let r_1 be the path between W' and Y', which cannot contain any $\mathbf{Z} \cup \mathbf{X}$ by condition (i) and definition of proper causal path. Also let r_2 be the path between W' and S which does not contain any $\mathbf{Z}$, and in particular, any $\mathbf{Z^T}; \mathbf{Z^S}$. Then, by Claim 4 condition (iii) is violated.
  - If the path between L' and W' has edges outgoing from W', then $Z \in \mathbf{Z} \cup S$ is a descendant of W'. And as in the previous argument either condition (i) or condition (iii) is violated.

- p needs to be blocked to satisfy condition (ii), which is possible only if W^0 is an inactive collider that blocks the path. However, this is not the case:
  - If $W^0 = L^0$, then W^0 is active because it is an ancestor of Z^0.
  - If W^0 is a collider in p, consider if it is connected to L^0 with an incoming or outgoing edge:
    * If the edge is incoming to W^0, it is also a collider in p_1 and has to be active by assumption.
    * If the edge is going out from W^0, then W^0 is an ancestor of Z and is active.

  Therefore, p is active and a contradiction is reached.

□

Proof. (of theorem 9). (If) Suppose the set $\mathbf{Z} = \mathbf{Z}^{\mathbf{T}} \cup \mathbf{Z}^{\mathbf{M}}$ satisfies the conditions of the criterion relative to the pair $\mathbf{X}$ and $\mathbf{Y}$ in a given causal diagram G.

Using the independences just proved in lemma 22 we proceed with a derivation of the target causal effect ending with the proposed adjustment expression:

We start the derivation by conditioning on $\mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}}, \mathbf{Z}^{\mathbf{T}}_{\mathbf{nd}}$ and L_1

$$P(\mathrm{y} \mid do(\mathrm{x})) \tag{F.15}$$

$$= \sum_{\mathbf{L}_1, \mathbf{Z}^{\mathbf{T}}_{\mathbf{nd}}, \mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}}} P(\mathrm{y} \mid do(\mathrm{x}), \mathrm{l}_1, \mathrm{z}^{\mathbf{T}}_{\mathbf{nd}}, \mathrm{z}^{\mathbf{S}}_{\mathbf{nd}}) P(\mathrm{l}_1, \mathrm{z}^{\mathbf{T}}_{\mathbf{nd}}, \mathrm{z}^{\mathbf{S}}_{\mathbf{nd}} \mid do(\mathrm{x})) \tag{F.16}$$

Since all the variables in the second term are non-descendants of $\mathbf{X}$ by definition, it holds that $(\mathbf{L}_1; \mathbf{Z}^{\mathbf{T}}_{\mathbf{nd}}; \mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}} \perp\!\!\!\perp \mathbf{X})_{G_{\overline{\mathbf{X}}}}$ and the third rule of the do-calculus can be applied to drop the $do()$ operator

$$= \sum_{\mathbf{L}_1;\mathbf{Z}^{\mathbf{T}}_{\mathbf{nd}};\mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{l}_1; \mathbf{z}^{\mathbf{T}}_{\mathbf{nd}}; \mathbf{z}^{\mathbf{S}}_{\mathbf{nd}}) P(\mathbf{l}_1; \mathbf{z}^{\mathbf{T}}_{\mathbf{nd}}; \mathbf{z}^{\mathbf{S}}_{\mathbf{nd}}) \tag{F.17}$$

We can employ independence (F.1) from lemma 22, $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}^{\mathbf{T}}_{\mathbf{d}}; \mathbf{Z}^{\mathbf{S}}_{\mathbf{d}} \mid \mathbf{L}_1; \mathbf{Z}^{\mathbf{T}}_{\mathbf{nd}}; \mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}}; \mathbf{X})_{G_{\overline{\mathbf{X}}}}$ to introduce the variables $\mathbf{Z}^{\mathbf{T}}_{\mathbf{d}}; \mathbf{Z}^{\mathbf{S}}_{\mathbf{d}}$ in the first term, after summing over the same variables in the second term

$$= \sum_{\mathbf{L}_1;\mathbf{Z}^{\mathbf{T}};\mathbf{Z}^{\mathbf{S}}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{l}_1; \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) P(\mathbf{l}_1; \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) \tag{F.18}$$

Applying the chain rule on the second term yields

$$= \sum_{\mathbf{L}_1;\mathbf{Z}^{\mathbf{T}};\mathbf{Z}^{\mathbf{S}}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{l}_1; \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) P(\mathbf{l}_1 \mid \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) P(\mathbf{z}^{\mathbf{T}} \mid \mathbf{z}^{\mathbf{S}}) P(\mathbf{z}^{\mathbf{T}}) \tag{F.19}$$

By definition of $\mathbf{Z}^{\mathbf{S}}$, $(\mathbf{Z}^{\mathbf{S}} \perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}})$, allowing us to introduce the S variable into the third factor

$$= \sum_{\mathbf{L}_1;\mathbf{Z}^{\mathbf{T}};\mathbf{Z}^{\mathbf{S}}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{l}_1; \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) P(\mathbf{l}_1 \mid \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) P(\mathbf{z}^{\mathbf{S}} \mid \mathbf{z}^{\mathbf{T}}; S{=}1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.20}$$

From lemma 22-(F.4), we use $(\mathbf{L}_1 \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}^{\mathbf{T}}; \mathbf{Z}^{\mathbf{S}})_{G_{\overline{\mathbf{X}(\mathbf{Z}^{\mathbf{T}};\mathbf{Z}^{\mathbf{S}})}}}$ to introduce the $do()$ operator into the second factor

$$= \sum_{\mathbf{L}_1;\mathbf{Z}^{\mathbf{T}};\mathbf{Z}^{\mathbf{S}}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{l}_1; \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) P(\mathbf{l}_1 \mid do(\mathbf{x}); \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}) P(\mathbf{z}^{\mathbf{S}} \mid \mathbf{z}^{\mathbf{T}}; S{=}1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.21}$$

Using the chain rule to combine the first and second factors. Sum out $\mathbf{L_1}$

$$= \sum_{\mathbf{Z^T};\mathbf{Z^S}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z^T}; \mathbf{z^S}) P(\mathbf{z^S} \mid \mathbf{z^T}; S=1) P(\mathbf{z^T}) \tag{F.22}$$

Using lemma 22-(F.3), $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z^T}; \mathbf{Z^S}; \mathbf{X})_{G_{\overline{\mathbf{X}}}}$, one can introduce the S variable into the first term

$$= \sum_{\mathbf{Z^T};\mathbf{Z^S}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z^T}; \mathbf{z^S}; S=1) P(\mathbf{z^S} \mid \mathbf{z^T}; S=1) P(\mathbf{z^T}) \tag{F.23}$$

Conditioning on $\mathbf{L_2}; \mathbf{Z^X}$ we get

$$= \sum_{\mathbf{Z^T};\mathbf{Z^S};\mathbf{L_2};\mathbf{Z^X}} P(\mathbf{y} | do(\mathbf{x}); \mathbf{z^T}; \mathbf{z^S}; \mathbf{l_2}; \mathbf{z^X}; S=1) P(\mathbf{l_2}; \mathbf{z^X} | do(\mathbf{x}); \mathbf{z^T}; \mathbf{z^S}; S=1) P(\mathbf{z^S} | \mathbf{z^T}; S=1) P(\mathbf{z^T}) \tag{F.24}$$

Using the independence $(\mathbf{L_2} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z^T}; \mathbf{Z^S}; S)_{G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}}$ from the definition of $\mathbf{L_2}$, and $(\mathbf{Z^X} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z^T}; \mathbf{Z^S}; S)_{G_{\overline{\mathbf{X}(\mathbf{Z^T};\mathbf{Z^S};S)}}}$ from the definition of $\mathbf{Z^X}$, we can remove the $do()$ operator from the second factor by applying rule 3 of do-calculus

$$= \sum_{\mathbf{Z^T};\mathbf{Z^S};\mathbf{L_2};\mathbf{Z^X}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z^T}; \mathbf{z^S}; \mathbf{l_2}; \mathbf{z^X}; S=1) P(\mathbf{l_2}; \mathbf{z^X} \mid \mathbf{z^T}; \mathbf{z^S}; S=1) P(\mathbf{z^S} \mid \mathbf{z^T}; S=1) P(\mathbf{z^T}) \tag{F.25}$$

By independence $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z^Y} \mid \mathbf{Z^T}; \mathbf{Z^S}; \mathbf{L_2}; \mathbf{Z^X}; \mathbf{X}; S)_{G_{\overline{\mathbf{X}}}}$ from lemma 22-(F.5), we can sum over $\mathbf{Z^Y}$ in the second term, move the new to the left and add $\mathbf{Z^Y}$ in the first term

$$= \sum_{\mathbf{Z};\mathbf{L_2}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}; \mathbf{l_2}; S=1) P(\mathbf{l_2}; \mathbf{z^Y}; \mathbf{z^X} \mid \mathbf{z^S}; \mathbf{z^T}; S=1) P(\mathbf{z^S} \mid \mathbf{z^T}; S=1) P(\mathbf{z^T}) \tag{F.26}$$

Rearranging using the chain rule

$$= \sum_{\mathbf{Z};\mathbf{L_2}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}; \mathbf{l_2}; S=1) P(\mathbf{l_2} \mid \mathbf{z}^{\mathbf{Y}}; \mathbf{z}^{\mathbf{X}}; \mathbf{z}^{\mathbf{T}}; \mathbf{z}^{\mathbf{S}}; S=1) P(\mathbf{z}^{\mathbf{Y}}; \mathbf{z}^{\mathbf{X}}; \mathbf{z}^{\mathbf{S}} \mid \mathbf{z}^{\mathbf{T}}; S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.27}$$

We can introduce $do(\mathbf{x})$ in the second term using the independence $(\mathbf{L_2} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}; S)_{G_{\overline{\mathbf{X}(\mathbf{Z};S)}}}$ from lemma 22-(F.6). Also considering that $\mathbf{Z} = \mathbf{Z}^{\mathbf{Y}} \cup \mathbf{Z}^{\mathbf{X}} \cup \mathbf{Z}^{\mathbf{T}} \cup \mathbf{Z}^{\mathbf{S}}$, we can rewrite as

$$= \sum_{\mathbf{Z};\mathbf{L_2}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}; \mathbf{l_2}; S=1) P(\mathbf{l_2} \mid do(\mathbf{x}); \mathbf{z}; S=1) P(\mathbf{z}^{\mathbf{Y}}; \mathbf{z}^{\mathbf{X}}; \mathbf{z}^{\mathbf{S}} \mid \mathbf{z}^{\mathbf{T}}; S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.28}$$

The first and second term can be combined using the chain rule. Then summing out L_2:

$$= \sum_{\mathbf{Z}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}; S=1) P(\mathbf{z}^{\mathbf{Y}}; \mathbf{z}^{\mathbf{X}}; \mathbf{z}^{\mathbf{S}} \mid \mathbf{z}^{\mathbf{T}}; S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.29}$$

Renaming the sets $\mathbf{Z}^{\mathbf{Y}} \cup \mathbf{Z}^{\mathbf{X}} \cup \mathbf{Z}^{\mathbf{S}}$ as $\mathbf{Z} \setminus \mathbf{Z}^{\mathbf{T}}$

$$= \sum_{\mathbf{Z}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}; S=1) P(\mathbf{z} \setminus \mathbf{z}^{\mathbf{T}} \mid \mathbf{z}^{\mathbf{T}}; S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.30}$$

From condition (ii) we have that $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}; S)_{G_{\underline{\mathbf{X}}}}$, then the $do()$ operator can be removed in the first term

$$= \sum_{\mathbf{Z}} P(\mathbf{y} \mid \mathbf{x}; \mathbf{z}; S=1) P(\mathbf{z} \setminus \mathbf{z}^{\mathbf{T}} \mid \mathbf{z}^{\mathbf{T}}; S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.31}$$

Since the adjustment holds in the magnified graph using only variables present in G, the same adjustment is admissible for the original model as well.

(Only if) For this direction of the proof we will establish that if the adjustment is valid then the conditions must be satisfied. In order to do so, we prove the contrapositive

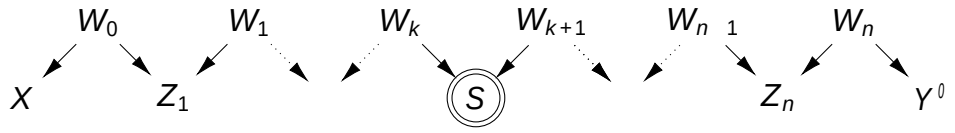


Figure F.1: Non-causal path between $\mathbf{X}$ and $\mathbf{Y}$ activated when S and $\mathbf{Z}$ is observed

statement, that is: failing to satisfy any of the conditions implies that the adjustment is not valid. First, let condition (ii)′ be a part of condition (ii) that says that all non-causal paths must be blocked given $\mathbf{Z}$ (without S). Then, (ii)′ will correspond to the second condition in the adjustment criterion [89]. First, assume that conditions (i) or (ii)′ do not hold. Then, the adjustment formula itself will not always identify the causal effect $P(\mathrm{y} \mid do(\mathrm{x}))$. For instance, consider any model compatible with $G_{\overline{S}}$ (which is also compatible with G). Then, the adjustment formula (6.6) reduces to adjustment without selection bias:

$$\sum_{\mathbf{z}} P(\mathrm{y} \mid \mathrm{x}, \mathbf{z}, S=1) P(\mathbf{z} \cap \mathbf{z}^{\mathbf{T}} \mid \mathbf{z}^{\mathbf{T}}, S=1) P(\mathbf{z}^{\mathbf{T}}) = \sum_{\mathbf{z}} P(\mathrm{y} \mid \mathrm{x}, \mathbf{z}) P(\mathbf{z}) \neq P(\mathrm{y} \mid do(\mathrm{x})) \tag{F.32}$$

The last inequality follows by the adjustment criterion [89], which implies that this expression will not always be equal to $P(\mathrm{y} \mid do(\mathrm{x}))$ whenever (i) or (ii)′ are not satisfied.

Now suppose conditions (i) and (ii)′ are satisfied but condition (ii) is not. Then, there exists a non-causal path p that is blocked given $\mathbf{Z}$ but is opened when S is observed. Path p must contain S as a collider and has the form $X' \dots \rightarrow S \leftarrow \dots Y'$ where $X' \in \mathbf{X}$ and $Y' \in \mathbf{Y}$. We need to find a model M compatible with a graph G that contains a path like p and show that the causal effect $P(\mathrm{y} \mid do(\mathrm{x}))$ is different from the adjustment expression (6.6). Consider a model compatible with the causal diagram depicted in fig. F.1.

The diagram evidences a non-causal path that is active when $\mathbf{Z} = \{Z_1, \dots, Z_n\}$ and S are observed. The elements in $\mathbf{Z}$ may be assigned to $\mathbf{Z}^{\mathbf{T}}$ in any way. Since the variables in $\mathbf{W} = \{W_0, \dots, W_n\}$ do not have any parents we can parametrize their distributions directly when constructing M. If we let every variable in $\mathbf{Z} \cup S$ behave as an `xor` of its parents,

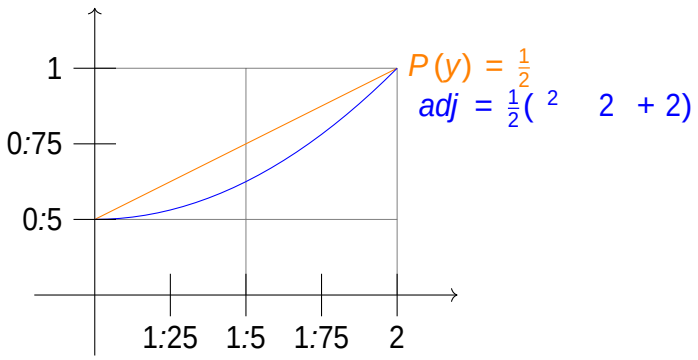


Figure F.2: Causal effect vs Adjustment expression for model M

$P(W_i{=}1) = \frac{1}{2}$; $i = f1;:::;ng$, where epsilon is a real constant, and $P(x) = P(w_0); P(y) = P(w_n)$ we obtain that:

$$P(Y{=}1 \mid do(X{=}1)) = P(Y{=}1) = \frac{1}{2}$$
$$\sum_z P(Y{=}1 \mid X{=}1; z; S{=}1)P(\mathbf{z} \cap \mathbf{z}^{\mathbf{T}} \mid S{=}1)P(\mathbf{z}^{\mathbf{T}}) = \frac{1}{2}(\quad 2 \quad + 2)$$

For any $\quad 2\ (1; 2)$ we have that the two quantities above are different (fig. F.2).

Then this is an M where the effect is not identifiable by the adjustment expression, because (ii)′ is not satisfied, impliying its necessity. If the path between any W_i and Z_j for any $0 \leq i \leq n; 1 \leq j \leq n; i \neq j$ has more variables in between we can make any variable be equal to its parent. Similarly, if Z_j is actually a descendant of a collider Q_j in the path, we can make every variable in the path to take the value of its parent, including Z_j. With those adjustments the model induces exactly the same distribution. If X or Y are parents of W_0 or W_n respectively the model is in the same equivalence class that the one presented and the conclusion holds.

Continuing with the remaining condition, assume that (i) and (ii) are satisfied but condition (iii) is not. Then, there exists a path p between S and some $Y' \in \mathbf{Y}$ that is active in the graph $G^{pbd}_{\mathbf{XY}}$ given $\mathbf{Z}^{\mathbf{T}}$. Consider the family of SCMs compatible with $G^{pbd}_{\mathbf{XY}}$. By condition (ii) the independence $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}; S)_{G^{pbd}_{\mathbf{XY}}}$ holds because all paths between $\mathbf{X}$ and

$\mathbf{Y}$ in $G^{pbd}_{\mathbf{XY}}$ are non-causal. Then, the adjustment expression for any model in that family can be reduced as follows:

$$\sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, S=1) P(\mathbf{z} \cap \mathbf{z}^{\mathbf{T}} \mid S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.33}$$

$$= \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{z}, S=1) P(\mathbf{z} \cap \mathbf{z}^{\mathbf{T}} \mid S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.34}$$

$$= \sum_{\mathbf{z}^{\mathbf{T}}} P(\mathbf{y} \mid \mathbf{z}^{\mathbf{T}}, S=1) P(\mathbf{z}^{\mathbf{T}}) \tag{F.35}$$

Equation (F.34) follows because of the independence $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}, S)_{G^{pbd}_{\mathbf{XY}}}$. The final expression is reached by combining the first two factors and summing out the variables in $\mathbf{Z} \cap \mathbf{Z}^{\mathbf{T}}$.

Since $G^{pbd}_{\mathbf{XY}}$ has no causal paths between $\mathbf{X}$ and $\mathbf{Y}$, the effect $P(\mathbf{y} \mid do(\mathbf{x})) = P(\mathbf{y})$. Also consider the effect $P(\mathbf{y} \mid do(s))$, which is always equals to $P(\mathbf{y})$. Consider the adjustment criterion relative to the pair $(S, \mathbf{Y})$ in $G^{pbd}_{\mathbf{XY}}$, note that $\mathbf{Z}^{\mathbf{T}}$ is not admissible since there is a non-causal path p that goes from S to $\mathbf{Y}$. Therefore, by the completeness of the adjustment criterion [89], there exists a model M compatible with $G^{pbd}_{\mathbf{XY}}$ where

$$P(\mathbf{y} \mid do(s)) \neq \sum_{\mathbf{z}^{\mathbf{T}}} P(\mathbf{y} \mid s, \mathbf{z}^{\mathbf{T}}) P(\mathbf{z}^{\mathbf{T}}) \tag{F.36}$$

The right hand side of eq. (F.36) includes eq. (F.35). We have then:

$$P(\mathbf{y} \mid do(\mathbf{x})) = P(\mathbf{y}) = P(\mathbf{y} \mid do(s)) \neq \sum_{\mathbf{z}^{\mathbf{T}}} P(\mathbf{y} \mid s, \mathbf{z}^{\mathbf{T}}) P(\mathbf{z}^{\mathbf{T}}) \tag{F.37}$$

Which proves that the adjustment expression does not rely the causal effect of interest in the model M, which is also compatible with G. □

Lemma 23 (Ancestral Path Separator)**.** *Let G be a causal diagram, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$*

be disjoint sets of variables in G. Let p be a path (not necessarily directed) with some $X' \in \mathbf{X}$ *and* $Y' \in \mathbf{Y}$ *as endpoints that is blocked (in the d-separation sense) when* $\mathbf{Z}$ *is observed. Let* $\mathbf{Z_A} = \mathbf{Z} \setminus An(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{W})$ *(i.e. the variables in* $\mathbf{Z}$ *that are ancestors any node in* $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{W}$*. Then p is also blocked when* $\mathbf{Z_A}$ *is observed.*

Proof. Assume for the sake of contradiction that *p* is active when $\mathbf{Z_A}$ is observed. This implies that there exists $Z' \in \mathbf{Z} \cap \mathbf{Z_A}$ that is needed to block *p*. For a variable Z' to block a path, at least one of the arrows in that path must be going out of Z' (i.e. Z' is not descendant of a collider in *p*). If we follow *p* starting at Z' in the direction of one of the outgoing arrows, there should be a collider before reaching X' (or Y' depending where the outgoing arrows is heading to in *p*), otherwise Z' would be an ancestor of X' ($\mathbf{Y}$) which is not the case by the definition of Z'. For *p* to be open, this collider must belong to $\mathbf{Z_A}$, but, since Z' is an ancestor of the collider Z' is also an ancestor of $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{W}$, contradicting its definition. Moreover, Z' cannot exists, neither *p*, and a contradiction is reached. □

Corollary 2 (Ancestral Separator Set)**.** *Let G be a causal diagram, and let* $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ *and* $\mathbf{W}$ *be disjoint sets of variables in G. Let* $\mathbf{Z_A} = \mathbf{Z} \setminus An(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{W})$*. If* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$ *in G, then* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_A})$ *in G.*

Proof. $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$ holds if and only if every path *p* with some $X' \in \mathbf{X}$ and $Y' \in \mathbf{Y}$ as endpoints, is blocked by $\mathbf{Z}$, by virtue of Theorem 23, all such paths are also blocked by $\mathbf{Z_A}$, implying $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_A})$. □

Proposition 9. *Suppose a pair* $\langle \mathbf{Z}, \mathbf{Z^T} \rangle$ *is admissible relative to* $\langle \mathbf{X}, \mathbf{Y} \rangle$ *in G. Then, the pair* $\langle \mathbf{Z_A}, \mathbf{Z_A^T} \rangle$*, where* $\mathbf{Z_A^T} = \mathbf{Z^T} \setminus An(\mathbf{X} \cup \mathbf{Y} \cup \{s\})$ *and* $\mathbf{Z_A} = \mathbf{Z} \setminus An(\mathbf{X} \cup \mathbf{Y} \cup \{s\})$*, is also admissible.*

Proof. Lets verify that $\mathbf{Z_A}$ satisfies each one of the four conditions of the criterion.

(a) Since $\mathbf{Z_A}$ is a subset of $\mathbf{Z}$ all its elements must satisfy this condition too.

(b) Any path between $\mathbf{X}$ and $\mathbf{Y}$ blocked by $\mathbf{Z}$, and in particular the non-causal ones, are also blocked by $\mathbf{Z_A}$ by virtue of lemma 23. If S is a descendant of a collider in some non-causal path p, it must be the case that $\mathbf{Z}$ blocks the subpath from between $\mathbf{X}$ and S or between S and $\mathbf{Y}$. Then again, by lemma 23, the set $\mathbf{Z_A}$ block the same subpath. Therefore, the overall path is blocked too.

(c) By corollary 2 $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}})_{G^{pbd}_{\mathbf{XY}}} =)\ \ (\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{Z}^{\mathbf{T}}_{\mathbf{A}})_{G^{pbd}_{\mathbf{XY}}}$.

Since all conditions hold, the criterion holds overall. □

Lemma 24 ([90])**.** *Let* $\mathbf{X}; \mathbf{Y}; \mathbf{I}; \mathbf{R}$ *be sets of nodes with* $\mathbf{I} \quad \mathbf{R}$, $\mathbf{R} \setminus (\mathbf{X} \cup \mathbf{Y}) = ;$. *If there exists an separator* $\mathbf{Z_0}$ *for* $\mathbf{X}; \mathbf{Y}$, *with* $\mathbf{I} \quad \mathbf{Z_0} \quad \mathbf{R}$ *then* $\mathbf{Z} = An(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I}) \setminus \mathbf{R}$ *is a separator for* $\mathbf{X}; \mathbf{Y}$.

Theorem 12 (Explicit admissible set construction)**.** *There exists an admissible pair in a causal diagram* G *relative to disjoint sets of variables* $\mathbf{X}; \mathbf{Y}$ *if and only if the pair* $\mathbf{Z}; \mathbf{Z}^{\mathbf{T}}$ *is admissible, where*

$$\mathbf{Z} = An(\mathbf{X} \cup \mathbf{Y} \cup fSg)_{G^{pbd}_{\mathbf{XY}}} \setminus \mathcal{C} \tag{6.15}$$

$$\mathbf{Z}^{\mathbf{T}} = (An(fSg \cup \mathbf{Y})_{G^{pbd}_{\mathbf{XY}}} \setminus \mathbf{T}) \setminus \mathcal{C} \tag{6.16}$$

Proof. This is easy to show using lemma 24. Suppose there exists some admissible pair $\mathbf{Z_0}; \mathbf{Z_0^T}$. theorem 9 implies that the pair must satisfy the conditions in definition 23. Furthermore, assume that $\mathbf{Z_0^T} \quad \mathbf{T}$ so that the adjustment is estimable from the assumed input. Then:

$$fSg \quad \mathbf{Z_0} \qquad\qquad \mathbf{C}; \tag{F.38}$$

$$; \quad \mathbf{Z_0^T} \qquad\qquad \mathbf{T} \setminus \mathbf{Z_0} \quad \mathbf{T} \setminus \mathbf{C} \tag{F.39}$$

$$\mathbf{Z_0} \;\; 2 \;\; \mathcal{Z}_{a;b}; \textit{and} \tag{F.40}$$

$$\mathbf{Z_0^T} \;\; 2 \;\; \mathcal{Z}_c \tag{F.41}$$

Applying lemma 24 to $\mathbf{Z_0}$ with $\mathbf{I} = fSg; \mathbf{R} = \mathbf{C}$ in graph $G^{pbd}_{\mathbf{XY}}$ we obtain the set (6.15). Using the same lemma on $\mathbf{Z_0^T}$ with $\mathbf{I} = ;; \mathbf{R} = \mathbf{T} \setminus \mathbf{C}$ in $G^{pbd}_{\mathbf{XY}}$ yields the set (6.16). And we have that:

$$\mathbf{Z} \;\; 2 \;\; \mathcal{Z}_{a;b}; \textit{and} \tag{F.42}$$

$$\mathbf{Z^T} \;\; 2 \;\; \mathcal{Z}_c \tag{F.43}$$

Therefore, the pair $\mathbf{Z}; \mathbf{Z^T}$ satisfied definition 23 which implies it is admissible. □

F.2 st-Adjustment Criterion

Theorem 15 (st-adjustment)**.** *Given a selection diagram G^{Δ} and disjoint sets of variables $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, the causal effect $P(\mathbf{y} \mid do(\mathbf{x}))$ is given by*

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid do(\mathbf{x}); \mathbf{z}; S=1) P(\mathbf{z}) \tag{6.33}$$

if and only if $\mathbf{Z}$ satisfies the st-adjustment criterion relative to $(\mathbf{X}; \mathbf{Y})$.

Proof. (if) Since $\mathbf{Z}$ satisfies condition (i), by theorem 14 we have,

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}) P(\mathbf{z}); \tag{F.44}$$

and by condition (ii) we have that $P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}) = P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}; S=1)$ therefore:

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}; S=1) P(\mathbf{z}): \tag{F.45}$$

(only if) To show the necessity of the criterion, we will prove that whenever its condi-

tions fail, there exists two models inducing G^{Δ} (i.e., $\langle M; M\ i$) that generate distributions distributions $P_{\mathbf{x}}(\mathbf{y}; \mathbf{z}\ j\ S=1)$ and $P\ (\mathbf{z})$ where $P_{\mathbf{x}}(\mathbf{y}) \neq \sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{y}\ j\ \mathbf{z})P\ (\mathbf{z})$.

If condition (i) is not satisfied, consider any two models $M; M'$ for which $P_{\mathbf{x}}(\mathbf{y}\ j\ \mathbf{z}) = P_{\mathbf{x}}(\mathbf{y}\ j\ \mathbf{z}; S=1)$, then by the e-adjustment criterion we have that the causal distribution is different that the g-adjustment.

Now, suppose condition (i) holds while condition (ii) does not. This is the case when there exists a path $\bar{p}$ from some T-node or S-node to $Y' \in \mathbf{Y}$ that is active given $\mathbf{Z}$ in $G^{\Delta}_{\overline{\mathbf{X}}}$.

First assume it is a T-node. Then, we will show that there exists a pair of models compatible with $G^{\Delta'}$, a graph where all edges but those in $\bar{p}$ have been removed, that is also compatible with G^{Δ}.

Note that $\bar{p}$ does not contain any variable in $\mathbf{X}$ otherwise it would be blocked in $G^{\Delta}_{\overline{\mathbf{X}}}$ given $\mathbf{X}$. Without loss of generality assume that Y' is the only element in $\mathbf{Y}$ also in $\bar{p}$, else select the other element in $\mathbf{Y}$ closer to T as Y'. Consequently, we have that

$$P_{\mathbf{x}}(\mathbf{y}) = P\ (\mathbf{y}) = P\ (\mathbf{y} \cap \{y'\})P\ (y') \tag{F.46}$$

Consider the possible structures of $\bar{p}$:

(a) Suppose that the path between T and Y' is directed: Assume that S is pointing to a node W at the beginning of the path (if T points directly to Y' we can use the same argument and sum-out W from the obtained distributions while substituting W with f_W in the equations). From eq. (F.46) we can sum over W and obtain

$$P_{\mathbf{x}}(\mathbf{y}) = P\ (\mathbf{y} \cap \{y'\}) \sum_{w} P\ (y'\ j\ w)P\ (w) \tag{F.47}$$

$$= P(\mathbf{y} \cap \{y'\}) \sum_{w} P(y'\ j\ w)P\ (w); \tag{F.48}$$

while the adjustment functional is equal to

$$\sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{z}) P(\mathbf{z}) \tag{F.49}$$

$$= \sum_{\mathbf{z}} P(\mathbf{y}) P(\mathbf{z}) \tag{F.50}$$

$$= P(\mathbf{y}) \tag{F.51}$$

$$= P(\mathbf{y} \cap \{y'\}) P(y') \tag{F.52}$$

$$= P(\mathbf{y} \cap \{y'\}) \sum_{w} P(y' \mid w) P(w). \tag{F.53}$$

We can parametrize two models with all variables binary, such that W is not independent of Y, and all conditional distributions are the same except for $P(w) \neq P(w)$. In this case the mapping between W and Y is one-to-one, hence the adjustment is different to the causal effect of interest.

(b) If the path $\bar{p}$ is not directed, suppose, as before, that T points to a variable W, which also has a parent R with binary domain, and let M be a model such that

$$P_{R=1}(y') \neq \sum_{\mathbf{z}} P(y' \mid \mathbf{z}; R = 1) P(\mathbf{z}); \tag{F.54}$$

which must exists by the completeness of the adjustment criterion [89] since $\mathbf{Z}$ does not satisfy it relative to $(R; Y')$. Now let M be a model such that $f_W(u_w) = f_W(u_w; R = 1)$. It is easy to see that $P_{R=1}(y') = P(y') = P(y' \mid \mathbf{x})$ and that $P(y' \mid \mathbf{z}; R = 1) = P(y' \mid \mathbf{z}) = P(y' \mid \mathbf{z}; \mathbf{x})$, hence equation (F.54) becomes

$$P_{\mathbf{x}}(y') \neq \sum_{\mathbf{z}} P(y' \mid \mathbf{z}; \mathbf{x}) P(\mathbf{z}); \tag{F.55}$$

implying that M ; M serve as counterexamples for this case.

Second, if the condition is violated because of an S-node, consider two identical sub-

models compatible compatible with $G^{\Delta'}$, a diagram where all edges but those witnessing the violation of the condition have been disconnected. Note that in $G^{\Delta'}$ no edges incoming to $\mathbf{X}$ remain and the transportability nodes are disconnected, therefore we have

$$P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}; S=1) = P'(\mathbf{y} \mid \mathbf{x}; \mathbf{z}; S=1); \tag{F.56}$$

because $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}; S=1)_{G^{\Delta'}_{\underline{\mathbf{X}}}}$. Then, the adjustment functional becomes

$$\sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{y}|\mathbf{z}; S=1)P'(\mathbf{z}) = \sum_{\mathbf{z}} P'(\mathbf{y}|\mathbf{x}; \mathbf{z}; S=1)P'(\mathbf{z}). \tag{F.57}$$

Since $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}; \mathbf{Z})_{G^{\Delta'}}$, theorem 9 implies that (F.57) is not equal to $P_{\mathbf{x}}(\mathbf{y})$ for some model compatible with $G^{\Delta'}$, which serves as a counter example to our case. □

F.2.1 Proof for Lemma 6

Lemma 6. *A set $\mathbf{Z}$ satisfies e-adjustment if and only if there exists $Z_i \in \mathbf{Z}$ such that Z_i satisfies eq.* (6.30) *or eq.* (6.31), *and $\mathbf{Z} \setminus \{Z_i\}$ satisfies e-adjustment.*

Proof. (if) Suppose $\mathbf{Z}$ satisfies e-adjustment, then there exists an order such that each element satisfies (6.30) or (6.31). Let Z_i be the last element in that order and notice that $\mathbf{Z} \setminus \{Z_1\}$ is also E-Admissible by the same order minus Z_i.

(only if) If $\mathbf{Z}$ satisfies e-adjustment and some Z_i satisfies (6.30) or (6.31) with $\mathbf{Z}^{i-1} = \mathbf{Z}$, then the order over $\mathbf{Z}$ with Z_i appended at the end witnesses that $\mathbf{Z} \cup \{Z_i\}$ satisfies e-adjustment. □

F.2.2 Proof for Lemma 7

Lemma 7. *If $\mathbf{Z}$ satisfies e-adjustment, then for any $Z_i \in \mathbf{Z}$ satisfying eq.* (6.30) *or eq.* (6.31), *the set $\mathbf{Z} \setminus \{Z_i\}$ satisfies e-adjustment.*

Proof. Define the predicates:

$E(\mathbf{Z})$: $\mathbf{Z}$ satisfies e-adjustment,

$\mathbf{A}(Z_i; \mathbf{Z})$: Z_i satisfies (6.30) with $\mathbf{Z}^{i-1} = \mathbf{Z} \setminus \{Z_1\}$, and

$\mathbf{B}(Z_i; \mathbf{Z})$: Z_i satisfies (6.31) with $\mathbf{Z}^{i-1} = \mathbf{Z} \setminus \{Z_1\}$.

Then from lemma 6 we have that

$$E(\mathbf{Z}) = \tag{F.58}$$

$$\forall_{Z_i \in \mathbf{Z}} (\mathbf{A}(Z_i; \mathbf{Z}) \text{ or } \mathbf{B}(Z_i; \mathbf{Z})) \text{ and } E(\mathbf{Z}): \tag{F.59}$$

Equivalently,

$$\neg(\exists_{Z_i \in \mathbf{Z}} (\neg \mathbf{A}(Z_i; \mathbf{Z}) \text{ and } \neg \mathbf{B}(Z_i; \mathbf{Z})) \text{ or } \neg E(\mathbf{Z} \setminus \{Z_i\})): \tag{F.60}$$

In particular, for any Z_i satisfying (6.30) or (6.31) this is implies

$$E(\mathbf{Z}) = \quad \neg(\mathit{false} \text{ and } \neg E(\mathbf{Z} \setminus \{Z_i\})) \tag{F.61}$$

$$E(\mathbf{Z}) = \quad E(\mathbf{Z} \setminus \{Z_i\}) \tag{F.62}$$

□

F.2.3 Proof for Theorem 16

Theorem 16. *$\mathbf{Z}$ satisfies e-adjustment (definition 26) w.r.t. $(\mathbf{X}; \mathbf{Y})$ in G if and only if IsEAdmissible (algorithm 12) returns true.*

Proof. (if) the procedure will return true only if it was able to remove all elements $\mathbf{Z} \setminus De(\mathbf{X})$ one by one while they satisfied either independence (6.30) or (6.31), witnessing that there exists an order satisfying the condition.

(only if) Suppose for the sake of contradiction that the algorithm returns false but $\mathbf{Z}$ satisfies e-adjustment. The procedure returns false when there is $\mathbf{Z}^{\emptyset} \subset \mathbf{Z}$ such that no

element in $\mathbf{Z}^0$ satisfies (6.30) or (6.31) and all elements in $\mathbf{Z} \setminus \mathbf{Z}^0$ removed in some order $Z_n^0 < \cdots < Z_1^0$ did (Z_n was the last removed before obtaining $\mathbf{Z}^0$). By lemma 7 we have that $\mathbf{Z}^0 \cup \{Z_n\}$ does not satisfy the criteria, and by repeatedly applying the same theorem we conclude that $\mathbf{Z}$ is not admissible, a contradiction. □