TECHNICAL REPORT
R-506
June 2021

UNIVERSITY OF CALIFORNIA

Los Angeles

Unit Selection Based on Counterfactual Logic

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Ang Li

2021

ABSTRACT OF THE DISSERTATION

Unit Selection Based on Counterfactual Logic

by

Ang Li
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2021
Professor Judea Pearl, Chair

The unit selection problem aims to identify a set of individuals who are most likely to exhibit a desired mode of behavior, which is defined in counterfactual terms. A typical example is that of selecting individuals who would respond one way if encouraged and a different way if not encouraged. Unlike previous works on this problem, which rely on ad-hoc heuristics, we approach this problem formally, using counterfactual logic, to properly capture the nature of the desired behavior. This formalism enables us to derive an informative selection criterion which integrates experimental and observational data. We show that a more accurate selection criterion can be achieved when structural information is available in the form of a causal diagram. We further discuss data availability issue regarding the derivation of the selection criterion without the observational or experimental data. We demonstrate the superiority of this criterion over A/B-test-based approaches.

The dissertation of Ang Li is approved.

Songchun Zhu

Guy Van den Broeck

Yizhou Sun

Judea Pearl, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This Ph.D. journey is the most memorable part of my life. First of all, it is hard to find words to express my gratitude to my advisor, Judea Pearl. You are the best advisor in this world. You led me to the realm of Causality. I enjoyed all the discussions with you during my Ph.D. Your gentle, passion, and patient guidance has taught me that what kind of researcher and person I should be. Your guidance is my greatest luck and wealth.

Thank you committee members Guy Van den Broeck, Yizhou Sun, and Songchun Zhu for your support and encouragement. All your advice will make me a better researcher.

I am proud that I am a member in Cognitive Systems Laboratory. When I was a new student at UCLA, Elias Bareinboim and Karthika Mohan helped me to get into the research and UCLA quickly. I also enjoyed all discussions with Chi Zhang and Scott Mueller. Parts of this work have benefited from discussions and collaborations with Scott Mueller. I enjoyed working with all of you and look forward to more collaborations in the future.

Thank you Kaoru Mulvihill for assistance with all administrative matters, and for all the help.

I would also like to thank my parents and my wife. Thank you my parents for all the support and encouragement since I was born. Thank you my wife, Ruirui Mao. We were together to the Unit States for our dream. We have been conquered lots of thorns in our life. Thank you for all your understanding and support, and wish you have a better student life in University of Wisconsin Madison.

Finally, thank you all my friends for your encouragement.

VITA

2010	B.S., Mathematics, Zhejiang University, Hangzhou, China.
2012	M.S., Computer Science, University of Minnesota Twin Cities, Minneapolis, MN, U.S.A.
2014-2021	Research Assistant, Cognitive Systems Lab, UCLA, Los Angeles, CA, U.S.A.

CHAPTER 1

Introduction

The problem of selecting individuals with a desired response pattern is encountered in many areas of industry, marketing, and health science. For example, in customer relationship management (CRM), it is of interest to predict which customers are about to churn but are likely to change their minds if enticed toward retention [BST99, Lej01, HYW06, TL09]. The cost associated with such programs compels the management to limit the enticements to customers who are most likely to exhibit the behavior of interest. In online advertising, as another example, companies are interested in identifying users who would click on an advertisement if and only if the advertisement is highlighted [YLW09, BPQ13, LCK14, SWY15]. The difficulty in identifying such users stems from the fact that the desired response pattern is not observed directly but rather is defined counterfactually in terms of what the individual would do under hypothetical unrealized conditions. For example, when we observe that a user has clicked on a highlighted advertisement, we do not know whether they would click on that same advertisement if it was not highlighted.

Individual behaviors are classified into four response types: complier, always-taker, never-taker, and defier [AIR96, BP97a]. Compliers are individuals who would respond positively if encouraged and negatively if not encouraged. Always-takers are individuals who always respond positively whether or not they are encouraged. Never-takers are individuals who always respond negatively whether or not they are encouraged. Defiers are individuals who would respond negatively if encouraged and positively if not encouraged.

A typical objective of the unit selection problem is to select individuals with the population-

specific characteristics that maximize the percentage of compliers because compliers represent the effectiveness of the encouragement.

A common solution that is explored in the literature is the A/B-test-based approach, where a controlled experiment is performed and the result is used as a criterion for selection. Specifically, users are randomly split into two groups called the control and treatment. Users in the control group are served unhighlighted advertisements, whereas those in the treatment group are served highlighted advertisements. Then, the population-specific characteristics that result in a higher difference between the two groups are used as predictors for the benefit of selection.

Departing from the prevailing literature, we treat the unit selection problem using the structural causal model (SCM) [Pea09], which accounts for the counterfactual nature of the desired behavior, and in which a large body of theoretical work has been established [GP98, Hal00].

The unit selection problem entails two subproblems, evaluation and search. The evaluation problem is to devise an estimable objective function that, if optimized over the set of population-specific characteristics C, would ensure an optimal counterfactual behavior for the selected group. The search task is to devise a search algorithm to select individuals (or population-specific characteristics) based on both their observed characteristics and the objective function devised above. This task is nontrivial due to the large number of characteristics available for each individual and the sparsity of data available in each cell (of characteristics).

Herein, we focus on the evaluation subproblem. In Chapter 4, we define a counterfactual expression that serves as the objective function for selection. This expression consists of probabilities of causation, such as the probability of necessity and sufficiency (PNS), which was studied in [Pea99, TP00, KC11].

Next, we provide two conditions under which the prevailing heuristic used in the literature

can become optimal. Our analysis shows that a selection criterion based on the A/B test heuristic can be made optimal (by fine tuning) under the condition of *monotonicity* or *gain equality*, which is defined formally in Chapter 5.2.

In general cases, however, counterfactual expressions are not identifiable. In Chapter 5.1, we derive tight bounds for this expression based on experimental and observational data and use the midpoint of the bounds as a selection criterion. Finally, through simulations, we demonstrate that sets of individuals selected by the derived criterion yield greater overall benefit than those selected using standard methods.

Recently, Mueller, Li, and Pearl [MLP21] proposed that information on covariates along with a causal structure could narrow the bounds of PNS. A similar technique could apply to our objective function. We then provide three graphical conditions under which the selection criterion could be improved.

CHAPTER 2

Motivating and Related Works

2.1 Motivating Example

Consider a mobile carrier that wants to identify customers likely to discontinue their services within the next quarter based on customer characteristics. The company management has access to user data, such as income, age, usage, and monthly payments. The carrier will then offer these customers a special renewal deal to dissuade them from discontinuing their services and to increase their service renewal rate. These offers provide considerable discounts to the customers, and the management prefers that the offers be made only to those customers who would continue their service if and only if they receive the offer. Note that some customers may discontinue service if and only if offered the renewal discount. Reasons for this could include being reminded that they are paying for a service they no longer want, feeling that discounts cheapen the service, reflecting on how much they are paying, being turned off by the promotional wording, or being annoyed by the process to claim the discount.

A typical aim is to select a subset of individuals with population-specific characteristics (we call it population-specific because the characteristics are only for categorizing individuals) c (a concrete instantiation of population-specific characteristics) that maximizes the percentage of compliers and minimizes that of defiers, always-takers, and never-takers among the selected customers (compliers are the customers who would continue the service if they received special offers and would not otherwise; defiers are the customers who would continue the service if they received no special offers and would not otherwise; always-takers are the

customers who would continue the service whether or not they received special offers; never-takers are the customers who would not continue the service whether or not they received special offers).

2.2 Related Works

There are two main approaches to the unit selection problem as extensively described in books, articles, and software packages.

The first approach relies on A/B testing and statistical analysis [SNO98, BCS01, Win01, RZS06, LR14]. Specifically, an experiment is conducted on a randomized controlled group of individuals. Then, the desired individuals (with concrete instantiation c of population-specific characteristics) are identified by maximizing the difference in the probability P(postive response$|c$, encouraged) $-$ P(positive response$|c$, not encouraged). However, the counterfactual nature of the desired behavior is not handled properly. A linear combination of P(postive response$|c$, encouraged) and P(positive response$|c$, not encouraged) does not maximize the percentage of compliers and minimize that of defiers, always-takers, and never-takers among the selected individuals. This is because the first term comprises compliers and always-takers and the second term comprises always-takers and defiers.

The second approach is based on machine-learning. Hung [HYW06] summarized and compared the most popular methods for churn prediction, including the regression, decision tree, and neural network methods. Using these approaches, a model is constructed using historical data to identify which customers are likely to discontinue their services. Then, the carrier offers a special renewal deal to the customers identified by the model as most likely to churn. However, an analysis of the set of customers who have accepted the special deal (hence, not churned) does not immediately reveal the customers who would have continued their services anyway and the customers who renewed their services only because of the special deal. Although another A/B test can be conducted, it leads to the same scenario as

that encountered when employing the above statistical approach.

The approach employed here differs fundamentally from those of previous studies by appealing to SCM, which is more robust and less prone to model misspecifications. First, the SCM model makes no assumptions about the data-generating process. Second, in most cases, experimental data can be evaluated in terms of observational data when a causal diagram is available. In such a case, observational data alone are sufficient for this approach. Third, and most importantly, the SCM properly accounts for the counterfactual nature of the desired behavior.

CHAPTER 3

Preliminaries

In this chapter, we review the counterfactual logic [GP98, Hal00, Pea09] associated with Pearl's SCM, which is used in the rest of this paper. Readers who are familiar with SCM may want to skip this chapter.

3.1 Do Calculus

We first review the do calculus, back-door and front-door criteria, and their associated adjustment formulas [Pea93, Pea95]. We use the causal diagrams in [Pea95, SGS00, Pea09, KF09].

The critical problem in causal analysis is to predict the results of interventions, such as those resulting from medical treatments or social programs, and this problem is denoted by $do(X = x)$ [Pea95]. We distinguish between cases in which a variable X takes a value x naturally and cases in which $X = x$ is fixed by denoting the latter as $do(X = x)$. Therefore, $P(Y = y|X = x)$ is the probability that $Y = y$ is conditional on the observation of $X = x$, and $P(Y = y|do(X = x))$ is the probability that $Y = y$ when we intervene to ensure that $X = x$. $P(Y = y|do(X = x))$ can be interpreted as experimental data.

A key concept of a causal diagram is d-separation [Pea14].

Definition 1 (d-separation)**.** *In a causal diagram G, a path p is blocked by a set of nodes 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 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 Z, and no descendant of B is in Z.*

If Z blocks every path between two nodes X and Y, then X and Y are d-separated conditional on Z, and thus are independent conditional on Z, denoted as $X \perp\!\!\!\perp Y \mid Z$.

With the concept of d-separation in a causal diagram, Pearl proposed the back-door and front-door criteria as follows:

Definition 2 (Back-door criterion)**.** *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the back-door criterion relative to (X, Y), if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

If a set of variables Z satisfies the back-door criterion for X and Y, the causal effects of X on Y are given by the adjustment formula:

$$P(y|do(x)) = \sum_z P(y|x, z)P(z). \tag{3.1}$$

Definition 3 (Front-door criterion)**.** *A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if:*

- *Z intercepts all directed paths from X to Y;*
- *there is no back-door path from X to Z; and*
- *all back-door paths from Z to Y are blocked by X.*

If a set of variables Z satisfies the front-door criterion for X and Y, and $P(x, Z) > 0$, then the causal effects of X on Y are given by the adjustment formula:

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x'). \tag{3.2}$$

The back-door and front-door criteria are powerful tools for estimating causal effects; however, causal effects are not identifiable if the set of adjustment variables Z is not fully observable. [TP00] provided the naivest bounds for causal effects (Equation 3.3), regardless of the causal diagram.

$$P(x, y) \leq P(y|do(x)) \leq 1 - P(x, y'). \tag{3.3}$$

3.2 Counterfactual Logic

The basic counterfactual statement associated with model M is denoted by $Y_x(u) = y$, and stands for: "Y would be y had X been x in unit $U = u$." Let M_x denote a modified version of M, with the equation(s) of set X replaced by $X = x$ (i.e., all edges that go into X have been removed). Then, the formal definition of the counterfactual $Y_x(u)$ is as follows:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{3.4}$$

In words, the counterfactual $Y_x(u)$ in model M is defined as the solution of Y in the "modified" submodel M_x. In [GP98, Hal00], a complete axiomatization of structural counterfactuals embracing both recursive and nonrecursive models is given.

Equation 3.4 implies that the distribution $P(u)$ induces a well-defined probability for the counterfactual event $Y_x = y$, written as $P(Y_x = y)$, which equals the probability that a random unit u would satisfy the equation $Y_x(u) = y$. Therefore, the probability of the event "Y would be y had X been x," $P(Y_x = y)$, is well-defined and $P(Y_x = y) = P(Y = y|do(X = x))$. $P(Y = y|do(X = x))$ can be interpreted as experimental data [Pea95]. With the same reasoning, the SCM assigns a probability to every counterfactual or combination of counterfactuals defined using the variables in SCM.

Using the above formal language for the counterfactual expression, all events involving a counterfactual scenario can be well defined because the event represented by the subscript does not actually occur. For example, $P(Y_x = y|X = x')$ defines the probability of the event

"Y would be y had X been x if we observed $X = x'$" (note that x and x' is a counterfactual scenario), $P(Y_x = y, Y_{x'} = y')$ defines the probability of the event "Y would be y had X been x and Y would be y' had X been x'" (note that x and x' is a counterfactual scenario, as well as y and y'), and $P(Y_x = y|X = x', Y = y')$ defines the probability of the event "Y would be y had X been x, if we observed $X = x'$ and $Y = y'$."

For simplicity, in the rest of the paper, we use y to denote the event $Y = y$, y' for the event $Y = y'$, x for the event $X = x$, x' for the event $X = x'$, y_x for the event $Y_x = y$, $y_{x'}$ for the event $Y_{x'} = y$, y'_x for the event $Y_x = y'$, and $y'_{x'}$ for the event $Y_{x'} = y'$.

CHAPTER 4

Counterfactual Formulation of Unit Selection Problem

Our objective is to find a set of population-specific characteristics c that maximizes the benefit associated with the resulting mixture of compliers, defiers, always-takers, and never-takers. Suppose the benefit of selecting a complier is β, the benefit of selecting an always-taker is γ, the benefit of selecting a never-taker is θ, and the benefit of selecting a defier is δ. We call $(\beta, \gamma, \theta, \delta)$ the benefit vector. Our objective, then, is to find c that maximizes the following expression:

$$argmax_c \ \beta P(\text{complier}|c) + \gamma P(\text{always-taker}|c) + \theta P(\text{never-taker}|c) + \delta P(\text{defier}|c).$$

Let $A = a$ denote that encouragement is received and $A = a'$ denote that encouragement is not received, and $R = r$ denote a positive response and $R = r'$ denote a negative response. Then, the objective function that maximizes the benefit on average of the selected individuals (the benefit function) can be formulated as follows:

$$argmax_c \ \beta P(r_a, r'_{a'}|c) + \gamma P(r_a, r_{a'}|c) + \theta P(r'_a, r'_{a'}|c) + \delta P(r'_a, r_{a'}|c). \tag{4.1}$$

Most importantly, this objective function can be bounded using observational and experimental data, as demonstrated in the following chapters.

CHAPTER 5

Selection Criterion without Causal Diagram

In this chapter, we derive an explicit solution to the unit selection problem using the benefit function with observational and experimental data.

5.1 General Selection Criterion

The first theorem (Theorem 4) is the most general theorem for evaluating the benefit function, which the only requirement is that population-specific characteristics C do not contain any descendant of a encouragement X.

Theorem 4. *The benefit function* $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ *is bounded as follows:*

$$\max\{p_1, p_2, p_3, p_4\} \leq f \leq \min\{p_5, p_6, p_7, p_8\} \text{ if } \sigma < 0,$$
$$\max\{p_5, p_6, p_7, p_8\} \leq f \leq \min\{p_1, p_2, p_3, p_4\} \text{ if } \sigma > 0,$$

where $\sigma, p_1, ..., p_8$ *are given by,*

$$\sigma = \beta - \gamma - \theta + \delta,$$

$$p_1 = (\beta - \theta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c),$$

$$p_2 = \gamma P(y_x|c) + \delta P(y'_x|c) + (\beta - \gamma)P(y'_{x'}|c),$$

$$p_3 = (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta + \delta)[P(y, x|c) + P(y', x'|c)],$$

$$p_4 = (\beta - \theta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c) +$$
$$+ (\beta - \gamma - \theta + \delta)[P(y, x'|c) + P(y', x|c)],$$

$$p_5 = (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c),$$

$$p_6 = (\beta - \theta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c),$$

$$p_7 = (\gamma - \delta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta + \delta)P(y|c),$$

$$p_8 = (\beta - \theta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c) - (\beta - \gamma - \theta + \delta)P(y|c).$$

Notably, the population-specific characteristics C cannot contain any descendant of X because if X is set to x and C contains a descendant of X, then C could be altered and $P(y_x|c)$ would be another unmeasurable counterfactual term.

The midpoint of the bounds in Theorem 4 is always used as the selection criterion. However, the lower (upper) bound can also be used, which can be interpreted as the average minimal (maximal) benefit gained from selecting one individual from the group.

5.2 Identifiability under Additional Assumptions

Herein, we show that the benefit function in Equation 4.1 can be evaluated precisely from pure experimental data under either of the conditions of monotonicity (Definition 5) and gain equality (Definition 7). Moreover, both conditions yield the same result.

5.2.1 Monotonicity

Monotonicity expresses the assumption that a change from $X =$ false to $X =$ true cannot, under any circumstance, change Y from true to false [TP00]. In epidemiology, this assumption is often expressed as "no prevention," that is, no individual in the population can be helped by exposure to the risk factor.

Definition 5. *(Monotonicity) A variable Y is said to be monotonic relative to variable X in a causal model M iff*

$$y'_x \wedge y_{x'} = \mathit{false}.$$

Theorem 6. *Given that Y is monotonic relative to X, the benefit function $f(c)$ is given by*

$$\begin{aligned} & f(c) \\ = \; & \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c) \\ = \; & (\beta - \theta)P(y_x|c) + (\gamma - \beta)P(y_{x'}|c) + \theta. \end{aligned}$$

5.2.2 Gain Equality

Gain equality states that the benefit of selecting a complier and a defier is the same as the benefit of selecting an always-taker and a never-taker (i.e., $\beta + \delta = \gamma + \theta$).

Definition 7. *(Gain equality) The benefit of selecting a complier (β), an always-taker (γ), a never-taker(θ), and a defier (δ) (benefit vector) is said to satisfy gain equality iff*

$$\beta + \delta = \gamma + \theta.$$

Theorem 8. *Given that the benefit vector $(\beta, \gamma, \theta, \delta)$ satisfies the gain equality, the benefit function $f(c)$ is given by*

$$\begin{aligned} & f(c) \\ = \; & \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c) \\ = \; & (\beta - \theta) P(y_x|c) + (\gamma - \beta) P(y_{x'}|c) + \theta. \end{aligned}$$

Note that for the special case where the perceived benefit is proportional to the final number of customers in the system, the A/B heuristic maximizes the expression $(H - D) \times P(y_x|c) - H \times P(y_{x'}|c)$, where H is the unit profit per remaining customer and D is the discount offered. This case corresponds to the following parameters in our notation $\beta = H - D, \gamma = -D, \theta = 0, \delta = -H$, therefore, Theorem 8 implies an identical benefict function $f = (H - D) \times P(y_x|c) - H \times P(y_{x'}|c)$. In other words, the A/B heuristic is optimal for this special case. For slightly more elaborate combinations of $(\beta, \gamma, \theta, \delta)$, however, Theorem 8 dictates a benefit function that is not captured by A/B heuristics.

Without either monotonicity or gain equality, we can only obtain bounds for the benefit function. However, in the next section, we demonstrate (by simulation) that taking the midpoint of the bounds as a selection criterion greatly improves the selection of individuals.

5.3 Examples

In this section, we present two simulated examples. One demonstrates that the midpoint of the bounds of the benefit function given by Theorem 4 are adequate for selecting the desired individuals, and the other demonstrates the case that satisfies the gain equality. In addition, we illustrate that, individuals selected using the traditional A/B-test-based statistical approach differ from those selected using the proposed approach, and they have a lower benefit on average.

5.3.1 Example in Churn Management

First, let us consider the motivating example in Chapter 2. Let $A = a$ denote the event that a customer receives the special deal, $A = a'$ denote the event that a customer receives no special deal, $R = r$ denote the event that a customer continues the services, $R = r'$ denote the event that a customer discontinues the services, and C (a set of variables) denote the population-specific characteristics of a customer (e.g., income, age, usage, and monthly payments). Figure 5.1 depicts the customer selection model.

The management estimates that the benefit of selecting a complier is \$100 as the profit is \$140 but the discount is \$40, the benefit of selecting an always-taker is −\$60 as the customer would continue the service anyway (so the company loses the value of the discount and an extra cost \$20 because the always-taker may require additional discounts in the future), the benefit of selecting a never-taker is \$0 as the cost of issuing the discount is negligible, and the benefit of selecting a defier is −\$140 as we lose a customer due to the special offer.

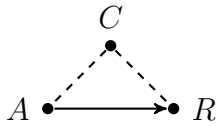

Figure 5.1: Causal diagram for the customer selection model.

Suppose they have two groups of customers, group 1 with characteristics c_1 and group 2 with characteristics c_2, and they have prior information that $P(r|c_1) = 0.7$ and $P(r|c_2) = 0.3$. They randomly select 700 customers from each group and offer the special renewal deal to 350 customers in each group. Table 5.1 summarizes the results.

Let us compare three selection strategies, each using a different objective function. The first is based on a simple A/B test heuristic, that is

$$\begin{aligned} Obj_1 &= argmax_c f_1(c) \\ &= argmax_c 100 \times P(r|c, do(a)) - 100 \times P(r|c, do(a')). \end{aligned}$$

Table 5.1: Results of a simulated study for churn management.

		$do(a)$	$do(a')$
Group 1	r	262	175
	r'	88	175
Group 2	r	87	52
	r'	263	298

Table 5.2: Results of three objective functions based on the data from the simulated study.

	f_1	f_2	f_3
Group 1	\$25	\$4.86	-\$2.63
Group 2	\$10	\$4.06	\$3.09

The second is based on a weighted A/B test heuristic approach, where

$$\begin{aligned} Obj_2 &= argmax_c f_2(c) \\ &= argmax_c 100 \times P(r|c, do(a)) - 140 \times P(r|c, do(a')). \end{aligned}$$

The third is based on the analysis of this study, where Equation 4.1 yields

$$\begin{aligned} Obj_3 &= argmax_c f_3(c) \\ &= argmax_c 100 \times P(r_a, r'_{a'}|c) + (-60) \times P(r_a, r_{a'}|c) + \\ &\quad +0 \times P(r'_a, r'_{a'}|c) + (-140) \times P(r'_a, r_{a'}|c). \end{aligned}$$

Then, we enter the data in Table 5.1 into the objective functions of groups 1 and 2. Table 5.2 summarizes the results (note that the midpoint of the bounds is used as the selection criterion for Obj_3 and $P(r_a|c) = P(r|c, do(a))$). The proposed approach selected group 2; however, the first and second objective functions selected group 1 as the desired individuals.

An informer with access to the fractions of compliers, always-takers, never-takers, and defiers in both groups (as summarized in Table 5.3, and these numbers are never known

Table 5.3: Percentages of four response types in each group for churn management.

	Complier	Always-taker	Never-taker	Defier
Group 1	30%	45%	20%	5%
Group 2	20%	5%	65%	10%

in reality) would easily conclude that the A/B-test-based approach had reached a wrong conclusion. In detail, the expected benefit of selecting an individual in group 1 is $100 \times 0.3 - 60 \times 0.45 + 0 \times 0.2 - 140 \times 0.05 = -\$4$, which means offering the special deal to group 1 would reduce the profit. The expected benefit of selecting an individual in group 2 is $100 \times 0.2 - 60 \times 0.05 + 0 \times 0.65 - 140 \times 0.1 = \$3$. Thus, the management should only offer the special deal to group 2.

Furthermore, Figure 5.2 depicts the benefit of group 1 from objective functions as a function of δ (β, γ, and θ are fixed), with each curve representing an objective function. The first two objective functions are the most common heuristics in the A/B-test-based approach. The third objective function is the real expected benefit. The last objective function is the midpoint of the bounds for the proposed objective function. We see that the midpoint of the bounds for the proposed objective function is the closest to the real benefit.

5.3.2 Example in Online Advertisement

5.3.2.1 Task 1

The management of a search engine company wants to decide whether it is worth sending an advertisement to a group of users, so as to maximize overall satisfaction. The management estimates that the satisfaction of recommending an advertisement to a complier is 2 degrees, as users would gain new information that they needed, that of recommending the advertisement to an always-taker is 1 degree, as users got a shortcut to the advertisement, that of recommending the advertisement to a never-taker is -1 degrees, as users got unnec-

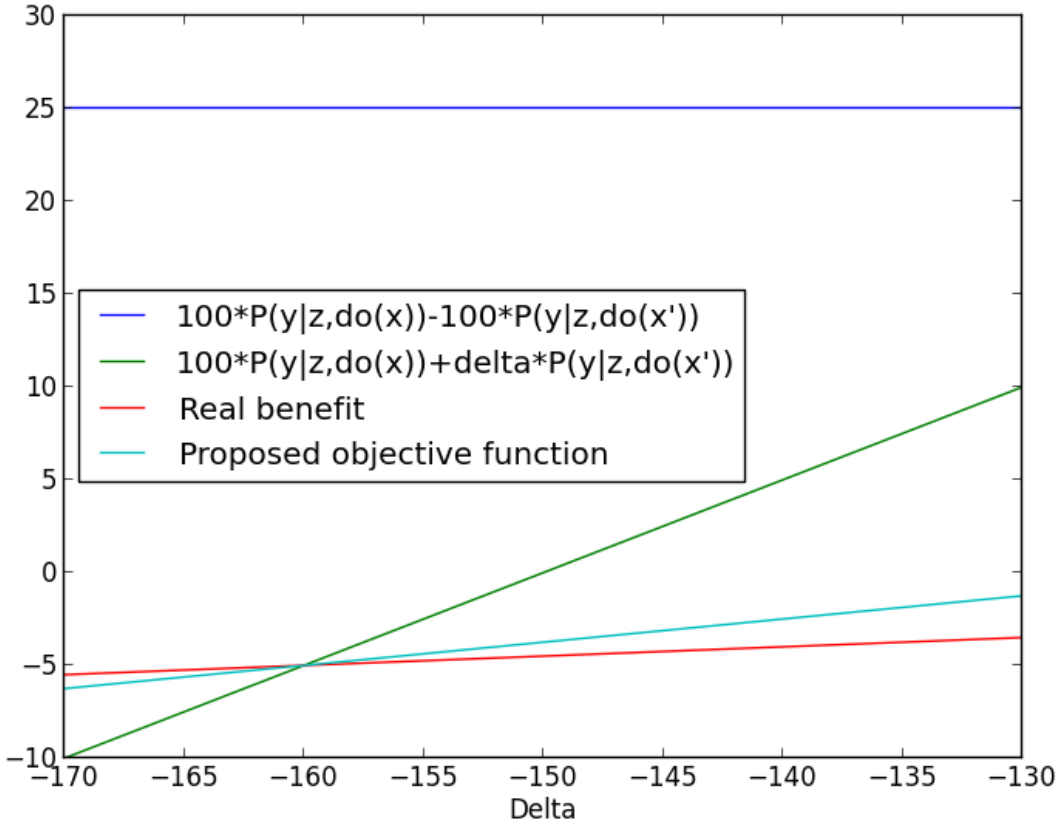

Figure 5.2: Benefit calculated from objective functions versus δ of group 1 in the churn management model.

Table 5.4: Results of a simulated study for advertisement recommendation.

	$do(a)$	$do(a')$
r	140	175
r'	210	175

essary information, and that of recommending the advertisement to a defier is -2 degrees, as the recommendation would prevent users to get needed information (compliers are the users who would click on the advertisement if the advertisement is recommended and would not if otherwise; always-takers are the users who would click on the advertisement whether or not the advertisement is recommended; never-takers are the users who would not click on the advertisement whether or not the advertisement is recommended; defiers are the users who would click on the advertisement if the advertisement is not recommended and would not if otherwise).

Let $A = a$ denote the event that the given advertisement is recommended, $A = a'$ denote the event that the given advertisement is not recommended, $R = r$ denote the event that a user clicks on the advertisement, and $R = r'$ denote the event that a user does not click on the advertisement.

Since no other data about the users are available, the management decides to conduct a randomized experiment and measure the degree to which the recommendation increases the users' click rate. The study involved 700 randomly selected users of whom 350 were recommended the advertisement. The results are listed in Table 5.4.

A simple A/B test heuristic approach concluded that recommending the advertisement to this group of users would increase the user satisfaction because
(satisfaction with recommendation) $\times P(r|do(a)) -$ (satisfaction without recommendation) $\times$ $P(r|do(a')) = 2 \times 0.4 - 1 \times 0.5 = 0.3$.

However, an informer with access to the fractions of compliers, always-takers, never-

Table 5.5: Percentages of four response types for advertisement recommendation.

Complier	Always-taker	Never-taker	Defier
30%	10%	20%	40%

takers, and defiers in the group (as summarized in Table 5.5, note that we will never know these numbers in reality because there is no monotonicity) claimed that the simple A/B test approach had reached the wrong conclusion. According to the company's assessment, the expected satisfaction per customer for recommending the advertisement to this group is $2 \times 0.3 + 1 \times 0.1 - 1 \times 0.2 - 2 \times 0.4 = -0.3$. This analysis shows that recommending the advertisement to this group of users would reduce the satisfaction. This is because only 30% of users are compliers and 10% are always-takers; therefore, a lot of advertisements are recommended to never-takers and defiers, which makes the recommendation reduce satisfaction.

In contrast, considering the benefit vector $(2, 1, -1, -2)$, we see that it satisfies the gain equality, which means that we can obtain the true average satisfaction even though we cannot determine the fraction of individuals in each response type. Accordingly, applying the benefit function of Theorem 8, we obtain that the expected satisfaction per user of recommending the advertisement to the group is $3 \times P(r|do(a)) - 1 \times P(r|do(a')) - 1 = 3 \times 0.4 - 1 \times 0.5 - 1 = -0.3$, which is precisely the satisfaction computed knowing the type distribution. This implies that the company should NOT recommend the advertisement to the group.

5.3.2.2 Task 2

Here, we consider two groups, c1 and c2. A study by the same company was conducted with 1400 randomly selected users (700 in each group), where the advertisement was recommended to 700 of those users (350 in each group). Table 5.6 summarizes the results.

A simple A/B test heuristic approach concluded that recommending the advertisement

Table 5.6: Results of a simulated study for advertisement recommendation with two groups.

		$do(a)$	$do(a')$
Group 1	r	140	88
	r'	210	262
Group 2	r	192	210
	r'	158	140

to both group of customers would increase the satisfaction because (satisfaction with recommendation)$\times P(r|do(a), c_1)-$(satisfaction without recommendation) $\times$ $P(r|do(a'), c_1) = 2 \times 0.4 - 1 \times 0.25 = 0.55$, and (satisfaction with recommendation) $\times$ $P(r|do(a), c_2)-$(satisfaction without recommendation)$\times P(r|do(a'), c_2) = 2\times 0.55-1\times 0.6 =$ 0.5.

However, an informer with access to the fractions of compliers, always-takers, never-takers, and defiers in both groups (as summarized in Table 5.7, note that we will never know these numbers in reality) claimed that the simple A/B test heuristic approach had reached a wrong conclusion. In detail, the expected satisfaction per user for recommending the advertisement to group 1 is $2\times 0.2+1\times 0.2-1\times 0.55-2\times 0.05 = -0.05$, which implies that recommending the advertisement to this group of users would reduce the satisfaction. The expected satisfaction per user for recommending the advertisement to group 2 is $2\times 0.3+1\times$ $0.25-1\times 0.1-2\times 0.35 = 0.05$, which implies that recommending the advertisement to this group of users would increase the satisfaction. Thus, the company should only recommend the advertisement to group 2.

In contrast, considering the benefit vector $(2, 1, -1, -2)$, we see that it satisfies the gain equality, which implies that we can obtain the true average satisfaction even though we cannot determine the fraction of individuals in each response type. Accordingly, applying the benefit function of Theorem 8, we obtain that the expected satisfaction per user for recommending the advertisement to the group 1 is $3\times P(r|do(a), c_1)-1\times P(r|do(a'), c_1)-1 =$

Table 5.7: Percentages of four response types in each group for advertisement recommendation.

	Complier	Always-taker	Never-taker	Defier
Group 1	20%	20%	55%	5%
Group 2	30%	25%	10%	35%

$3 \times 0.4 - 1 \times 0.25 - 1 = -0.05$, and the expected satisfaction per user for recommending the advertisement to the group 2 is $3 \times P(y|do(x), c_2) - 1 \times P(y|do(x'), c_2) - 1 = 3 \times 0.55 - 1 \times 0.6 - 1 = 0.05$, which is precisely the satisfaction computed knowing the type distribution. This implies that the company should NOT recommend the advertisement to group 1.

5.4 Discussion

In this section, we discuss additional features of the counterfactual-logic-based approach. First, as discussed in Chapter 4, the objective function properly accounts for the counterfactual nature of the desired behavior. Theorem 8 provides theoretical assurance that the A/B-test-based approach can be made optimal under certain conditions. However, the second simulated experimental example demonstrates that this approach selected individuals with a lower expected benefit when the cost-benefit structure is ignored. Although the proposed objective function is, in general, not identifiable and cannot be used in selection, the previous section shows that the midpoint of the tight bounds in Theorem 4 is adequate for selecting the desired individuals.

Second, considering a causal diagram and a set of observed variables that satisfies the back-door or front-door criterion (Definitions 2 and 3), Theorem 4 can be applied using purely observational data via the adjustment formula (Equations 3.1 and 3.2).

Third, the proposed approach could be used to evaluate machine learning models as well as to generate labels for machine learning models. The accuracy of such machine learning

models would be high because they consider counterfactual scenarios.

Fourth, Theorem 8 provides a way for identifying the weight coefficients in the extensively used statistical approach when the additional assumption is satisfied.

Finally, the proposed approach is applicable universally to any application in which the manager can assess the benefits associated with selecting a unit in each of the four response types of units (benefit vector). Theorem 4 ensures that for any benefit vector input, we obtain the desired output. The benefit vector input is not determined by the model but by the manager who can use the algorithm for any combination of inputs.

CHAPTER 6

Selection Criterion with Graphical Conditions

The unit selection problem discussed in Chapter 5 does not require the information of covariates, and the variables C in Theorems 4 and 8 are only population-specific variables and categorize populations. Therefore, Theorems 4 and 8 can be employed in any application, provided the population-specific variables C contain no descendant of X. Recently, Mueller, Li, and Pearl [MLP21] proposed that the information of covariates along with a causal structure could narrow the bounds of PNS. A similar technique could apply to the benefit function.

6.1 Motivating Example

Consider that a carwash company wants to offer a discount to Google employees. The manager of the company wants to maximize the total profit, including the nonimmediate profit. The management estimates that the benefit of selecting a complier is \$100 as the profit is \$140 but the discount is \$40, that of selecting an always-taker is $-$\$60 as the customer would use the service anyway (so the company loses the value of the discount and an extra cost \$20 because the always-taker may require additional discounts in the future), that of selecting a never-taker is \$0 as the cost of issuing the discount is negligible, and that of selecting a defier is $-$\$140 as a customer is lost due to the discount. The manager has both experimental and observational data from the Google company associated with the age of customers. The manager wants to know the average profit if they offer the discount to the Google employees.

Let $A = a$ denote the event that a customer receives the discount, $A = a'$ denote the event that a customer receives no discount, $R = r$ denote the event that a customer uses the services, $R = r'$ denote the event that a customer does not use the services, $C = c$ denote a Google customer, $C = c'$ denote a Facebook customer, $Z = z$ denote a younger costumer (age below or equal to 50), and $Z = z'$ denote a older costumer (age above 50). The model is as shown in Figure 6.1.

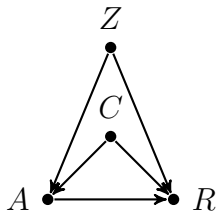

Figure 6.1: Company selection model.

The manager collected the data listed in Tables 6.1 and 6.2 from Google, and the benefit vector is $(100, -60, 0, -140)$. The benefit function can easily be applied to Equation 4.1 as follows:

$$argmax_c \; 100P(r_a, r'_{a'}|c) - 60P(r_a, r_{a'}|c) + 0P(r'_a, r'_{a'}|c) - 140P(r'_a, r_{a'}|c). \tag{6.1}$$

From Theorem 4, we enter the data in Tables 6.1 and 6.2 into the benefit function. The bounds of the benefit function are $[-0.377, 2.892]$, and the midpoint is 1.635. It suggests that the carwash company would gain \$1.635 profit from each individual from Google if they offer Google employees the discount. Besides, the majority of the bounded area is positive, which provided more confidence that the conclusion is correct. However, Theorem 4 only uses the overall data in Tables 6.1 and 6.2. We will illustrate later that considering the covariate (age), the bounds of the benefit function would reduce to $[-0.122, -0.046]$, and the midpoint is -0.084. It suggests that the carwash company would lose \$0.084 profit from each individual from Google if they offer Google employees the discount. Notably, the upper bound (-0.046) is negative, which means that the company must lose the profit.

Table 6.1: Experimental data collected by the carwash company.

	Discount	No Discount
Young	44.6% used the service	4.5% used the service
Elder	99.5% used the service	72.0% used the service
Overall	83.7% used the service	52.5% used the service

Table 6.2: Observational data collected by the carwash company.

	Discount	No Discount
Young	90 out of 152 used the service (59.2%)	9 out of 50 used the service (18%)
Elder	157 out of 159 used the service (98.7%)	239 out of 339 used the service (70.5%)
Overall	247 out of 311 used the service (79.4%)	248 out of 389 used the service (63.8%)

6.2 Selection Criteria with Causal Diagrams

6.2.1 Causal Diagram with Nondescendant Covariates

Theorem 9 provides bounds for the benefit function when a set Z of variables can be measured, which satisfy only one condition: both population-specific variables C and covariates Z contain no descendant of X. This condition is important because if X is set to x and $C \cup Z$ contains a descendant of X, then $C \cup Z$ could be altered and $P(y_x|z,c)$ would be another unmeasurable counterfactual term. If the descendant is independent of Y_x, then $P(y_x|z,c)$ would be measurable, but the descendant would not contribute to any narrowing of the bounds. These bounds are always contained within the bounds of the benefit function in Theorem 4.

Theorem 9. *Given a causal diagram G and distribution compatible with G, let $Z \cup C$ be a set of variables that does not contain any descendant of X in G, then the benefit function $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ is bounded as follows:*

$$W + \sigma U \leq f \leq W + \sigma L \qquad \text{if } \sigma < 0,$$
$$W + \sigma L \leq f \leq W + \sigma U \qquad \text{if } \sigma > 0,$$

where σ, W, L, U *are given by,*

$$\sigma = \beta - \gamma - \theta + \delta,$$

$$W = (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c),$$

$$L = \sum_z \max \left\{ \begin{array}{c} 0, \\ P(y_x|z,c) - P(y_{x'}|z,c), \\ P(y|z,c) - P(y_{x'}|z,c), \\ P(y_x|z,c) - P(y|z,c) \end{array} \right\} \times P(z|c),$$

$$U = \sum_z \min \left\{ \begin{array}{c} P(y_x|z,c), \\ P(y'_{x'}|z,c), \\ P(y,x|z,c) + P(y',x'|z,c), \\ P(y_x|z,c) - P(y_{x'}|z,c) + P(y,x'|z,c) + P(y',x|z,c) \end{array} \right\} \times P(z|c).$$

Notably, C can be interpreted as the population-specific variables, and Z are the contributions in each population.

Now, we consider the motivating example at the beginning of this chapter. From Theorem 9, we enter the data in Tables 6.1 and 6.2 into the benefit function, the bounds of the benefit function is $[-0.122, -0.046]$, where the midpoint is -0.084. This suggests that the carwash company would lose \$0.084 profit from each individual from Google if they offer Google employees the discount. Notably, the upper bound (-0.046) is negative, which means that the carwash company must lose the profit.

6.2.2 Causal Diagram with Mediators

6.2.2.1 Partial Mediators

In Figure 6.2, Z is a descendant of X; thus, we cannot use Theorem 9. However, the absence of confounders between Z and Y and between X and Y permits us to bound the benefit function as follows:

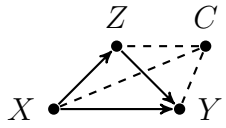

Figure 6.2: Mediator Z with direct effects.

Theorem 10. *Given a causal diagram G and distribution compatible with G, let Z be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp\!\!\!\perp X \cup Z_{x'} \mid Z_x, C)$ in G, and C does not contain any descendant of X in G, then the benefit function $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ is bounded as follows:*

$$W + \sigma U \leq f \leq W + \sigma L \qquad \text{if } \sigma < 0,$$
$$W + \sigma L \leq f \leq W + \sigma U \qquad \text{if } \sigma > 0,$$

where σ, W, L, U are given by,

$$\sigma = \beta - \gamma - \theta + \delta,$$
$$W = (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c),$$
$$L = \max \left\{ \begin{array}{c} 0, \\ P(y_x|c) - P(y_{x'}|c), \\ P(y|c) - P(y_{x'}|c), \\ P(y_x|c) - P(y|c) \end{array} \right\},$$
$$U = \min \left\{ \begin{array}{c} P(y_x|c), \\ P(y'_{x'}|c), \\ P(y, x|c) + P(y', x'|c), \\ P(y_x|c) - P(y_{x'}|c) + P(y, x'|c) + P(y', x|c), \\ \sum_z \sum_{z'} \min\{P(y|z, x, c), P(y'|z', x', c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\} \end{array} \right\}.$$

Note that although this lower bound is unchanged from that in Theorem 4, the upper bound contains a vital additional argument to the min function. This new term can significantly reduce the upper bound. The rest of the terms are included because sometimes the

bounds of Theorem 4 are superior. The following theorem has the same quality.

6.2.2.2 Pure Mediators

Figure 6.3 is a special case of Figure 6.2, in which X has no direct effects on Y. The resulting bounds for the benefit function are as follows:

Theorem 11. *Given a causal diagram G in Figure 6.3 and distribution compatible with G, and C does not contain any descendant of X, then the benefit function $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ is bounded as follows:*

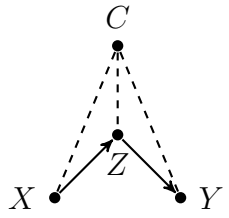

Figure 6.3: Mediators Z with no direct effects.

$$W + \sigma U \leq f \leq W + \sigma L \qquad \text{if } \sigma < 0,$$
$$W + \sigma L \leq f \leq W + \sigma U \qquad \text{if } \sigma > 0,$$

where σ, W, L, U *are given by,*

$$\sigma = \beta - \gamma - \theta + \delta,$$

$$W = (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c),$$

$$L = \max \left\{ \begin{array}{c} 0, \\ P(y_x|c) - P(y_{x'}|c), \\ P(y|c) - P(y_{x'}|c), \\ P(y_x|c) - P(y|c) \end{array} \right\},$$

$$U = \min \left\{ \begin{array}{c} P(y_x|c), \\ P(y'_{x'}|c), \\ P(y, x|c) + P(y', x'|c), \\ P(y_x|c) - P(y_{x'}|c) + P(y, x'|c) + P(y', x|c), \\ \Sigma_z \Sigma_{z' \neq z} \min\{P(y|z, c), P(y'|z', c)\} \times \min\{P(z|x, c), P(z'|x', c)\} \end{array} \right\}.$$

The core terms for Theorem 11 added to the upper bound notably only require observational data.

6.3 Simulation Study

In this section, we illustrate that the bounds of the benefit function are improved by Theorem 9 in a simple causal diagram, as shown in Figure 6.4.

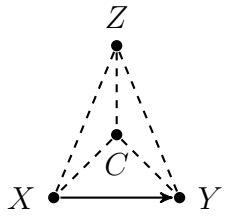

Figure 6.4: Causal diagram such that $C \cup Z$ is not a descendant of X.

We randomly generated 100000 sample distributions (observational data and experimental data) compatible with the causal diagram by Algorithm 1 (X, Y, Z are binary). The

algorithm ensures that the experimental data satisfies the general relation with the observational data in Equation 3.3. We set the benefit vector $(\beta, \gamma, \theta, \delta)$ to be the most common one $(1, -1, -1, -1)$ with the aim to encourage compliers while avoiding always-takers, never-takers, and defiers. For sample distribution i, let $[a_i, b_i]$ be the bounds with the causal diagram from Theorem 9 and $[c_i, d_i]$ be the bounds without the causal diagram from Theorem 4. We summarized the following criteria:

- Average increased lower bound : $\frac{\sum(a_i - c_i)}{100000}$;
- Average decreased upper bound : $\frac{\sum(d_i - b_i)}{100000}$;
- Average gap without the causal diagram : $\frac{\sum(d_i - c_i)}{100000}$;
- Average gap with the causal diagram : $\frac{\sum(b_i - a_i)}{100000}$;
- Number of sample distributions in which the decision was flipped : $\sum e_i$ where, $e_i = 1$ if $(a_i + b_i) \times (c_i + d_i) < 0$ and $e_i = 0$ otherwise;
- Number of sample distributions in which the bounds with the causal diagram from Theorem 9 were narrower : $\sum f_i$ where, $f_i = 1$ if $(a_i \neq c_i)$ *or* $(b_i \neq d_i)$ and $f_i = 0$ otherwise.

The results are summarized in Table 6.3. We then randomly picked 100 out of 100000 sample distributions to draw the graph of bounds with and without the causal diagram. The results are shown in Figure 6.5.

Algorithm 1: Generate sample distributions compatible with the causal diagrams 6.4 and 7.1.

input : n, number of sample distributions needed.

output: n sample distributions (observational data and experimental data).

begin

  for $i \leftarrow 1$ **to** n **do**

    $t_1 \leftarrow$ `random-uniform(0,1)` $\times 1000$;

    $t_2 \leftarrow$ `random-uniform(0,1)` $\times (1000 - t_1)$;

    $t_3 \leftarrow$ `random-uniform(0,1)` $\times (1000 - t_1 - t_2)$;

    $t_4 \leftarrow 1000 - t_1 - t_2 - t_3$;

    $o_1 \leftarrow$ `random-uniform(0,1)` $\times t_1$;

    $o_2 \leftarrow$ `random-uniform(0,1)` $\times t_2$;

    $o_3 \leftarrow$ `random-uniform(0,1)` $\times t_3$;

    $o_4 \leftarrow$ `random-uniform(0,1)` $\times t_4$;

    $P(y|do(x), z)[i] \leftarrow$ `random-uniform(0,1)` $\times \frac{t_2}{t_1+t_2} + \frac{o_1}{t_1+t_2}$;

    $P(y|do(x'), z)[i] \leftarrow$ `random-uniform(0,1)` $\times \frac{t_1}{t_1+t_2} + \frac{o_2}{t_1+t_2}$;

    $P(y|do(x), z')[i] \leftarrow$ `random-uniform(0,1)` $\times \frac{t_4}{t_3+t_4} + \frac{o_3}{t_3+t_4}$;

    $P(y|do(x'), z')[i] \leftarrow$ `random-uniform(0,1)` $\times \frac{t_3}{t_3+t_4} + \frac{o_4}{t_3+t_4}$;

    $P(x, y, z)[i] \leftarrow o_1/1000$;

    $P(x, y, z')[i] \leftarrow o_3/1000$;

    $P(x, y', z)[i] \leftarrow (t_1 - o_1)/1000$;

    $P(x, y', z')[i] \leftarrow (t_3 - o_3)/1000$;

    $P(x', y, z)[i] \leftarrow o_2/1000$;

    $P(x', y, z')[i] \leftarrow o_4/1000$;

    $P(x', y', z)[i] \leftarrow (t_2 - o_2)/1000$;

    $P(x', y', z')[i] \leftarrow (t_4 - o_4)/1000$;

  end

end

Table 6.3: Simulation results of 100000 sample distributions compatible with the causal diagram in Figure 6.4.

Average increased lower bound	Average decreased upper bound	Average gap by Theorem 4	Average gap by Theorem 9	Decision flipped	Bounds narrower
0.0494	0.0496	0.4342	0.3352	920	93688

From Figure 6.5, we can see that the bounds of the benefit function are improved in most of the samples with the causal diagram. We can see in Table 6.3 that the average gap without the causal diagram is 0.4342, while the average gap with the causal diagram is 0.3352, and both the lower bound and upper bound are improved by roughly 0.05. The decisions flipped (i.e., the results of Theorem 4 suggest gain profit, while the results of Theorem 9 suggest lose profit, or the reverse) are $920/100000 \approx 1\%$ of the samples, which means that at least 1% of the applications would have the wrong decision if we do not consider covariates. The bounds with the causal diagram are actually narrower in $93688/100000 \approx 93.7\%$ of the samples. Therefore, if a set of Z is available that satisfies Theorem 9, the bounds of the benefit function are more useful as the gap is narrower.

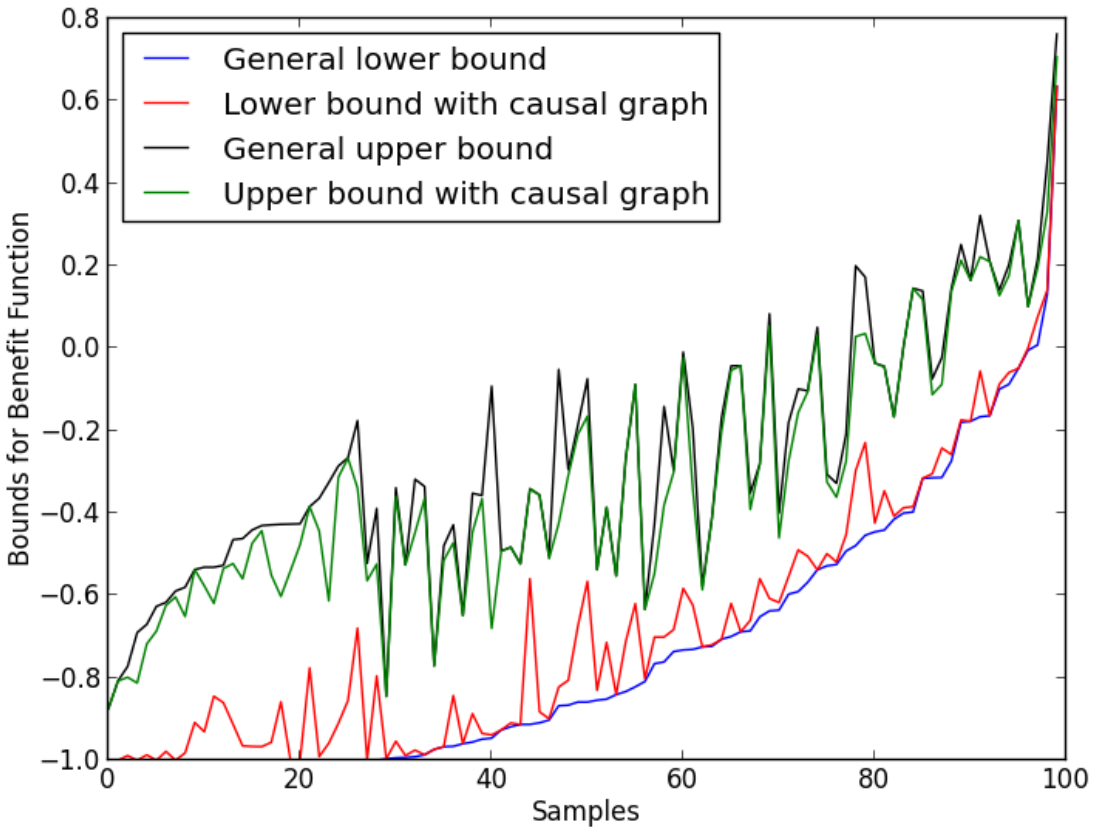

Figure 6.5: Bounds of the benefit function for 100 samples in the causal diagram of Figure 6.4, where the general bounds are obtained from Theorem 4 and the bounds with the causal diagram are obtained from Theorem 9.

CHAPTER 7

Data Availability

In previous chapters, we showed how to bound the benefit function using observational data and experimental data. However, in reality, we may not have both data. In this chapter, we will illustrate how to evaluate the benefit function with only experimental or observational data.

7.1 Unit Selection with Experimental Data

Consider the situation that only experimental data are available to us. This is very common in reality, for example, a new drug that was never developed in the world and only went through an experimental study. It is not easy to estimate observational data from the experimental data. Therefore, the first step is to check if the benefit vector $(\beta, \gamma, \theta, \delta)$ satisfies the gain equality or if the outcome is monotonic relative to the encouragement in Theorem 8; if so, the benefit function then becomes a point estimate and only depends on the experimental data. Otherwise, although the observational data is difficult to estimate from the experimental data, the experimental data play a major role in Theorem 4. Therefore, we can simply remove the observational data terms in the theorem and still have informative bounds for the benefit function. Theorem 4 then becomes as follows:

Theorem 12. *The benefit function* $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) +$

$\delta P(y_{x'}, y'_x|c)$ is bounded as follows:

$$\max\{p_1, p_2\} \leq f \leq \min\{p_3, p_4\} \text{ if } \sigma < 0,$$
$$\max\{p_3, p_4\} \leq f \leq \min\{p_1, p_2\} \text{ if } \sigma > 0,$$

where $\sigma, p_1, ..., p_4$ are given by,

$$\begin{aligned}
\sigma &= \beta - \gamma - \theta + \delta, \\
p_1 &= (\beta - \theta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
p_2 &= \gamma P(y_x|c) + \delta P(y'_x|c) + (\beta - \gamma)P(y'_{x'}|c), \\
p_3 &= (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
p_4 &= (\beta - \theta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c).
\end{aligned}$$

Notably, there is no experimental data only version of Theorems 9, 10, and 11 because covariates Z itself require observational data; at least $P(Z|C)$ is needed. Further, if we have observational data partially available, the corresponding observational terms in Theorem 4 could be added back to Theorem 12.

7.1.1 Simulation Study

In this section, we illustrate that the bounds of the benefit function given by Theorem 12 are still informative by simulation in a simple causal diagram as shown in Figure 7.1.

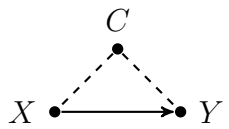

Figure 7.1: Simple causal diagram with population-specific variable C.

Similarly as that in Chapter 6.3, we randomly generated 100000 sample distributions (observational data and experimental data) compatible with the causal diagram in Figure 7.1

using Algorithm 1 (summing over Z with the outputs of the algorithm). We set the benefit vector $(\beta, \gamma, \theta, \delta)$ to be the most common one $(1, -1, -1, -1)$ with the aim to encourage compliers while avoiding always-takers, never-takers, and defiers. For sample distribution i, let $[a_i, b_i]$ be the bounds with the experimental data and observational data from Theorem 4 and $[c_i, d_i]$ be the bounds with the experimental data only from Theorem 12. We summarized the following criteria:

- Average decreased lower bound : $\frac{\sum(a_i - c_i)}{100000}$;
- Average increased upper bound : $\frac{\sum(d_i - b_i)}{100000}$;
- Average gap with the experimental data only : $\frac{\sum(d_i - c_i)}{100000}$;
- Average gap with the experimental data and observational data : $\frac{\sum(b_i - a_i)}{100000}$;
- Number of sample distributions in which the decision was flipped : $\sum e_i$ where, $e_i = 1$ if $(a_i + b_i) \times (c_i + d_i) < 0$ and $e_i = 0$ otherwise;
- Number of sample distributions in which the bounds with the experimental data only from Theorem 12 were wider : $\sum f_i$ where, $f_i = 1$ if $(a_i \neq c_i)$ *or* $(b_i \neq d_i)$ and $f_i = 0$ otherwise.

The results are summarized in Table 7.1. We then randomly picked 100 out of 100000 sample distributions to draw the graph of the bounds with and without the observational data. The results are shown in Figure 7.2.

From Figure 7.2, we can see that although the bounds of the benefit function are wider in most of the samples in the absence of observational data, the bounds with the experimental data only are still valid estimates of the benefit function. We can see in Table 7.1 that the average gap with the experimental data and observational data is 0.4346, while the average gap with the experimental data only becomes 0.5490; both the lower bound and upper bound are affected by roughly 0.06. The decisions flipped (i.e., the results from

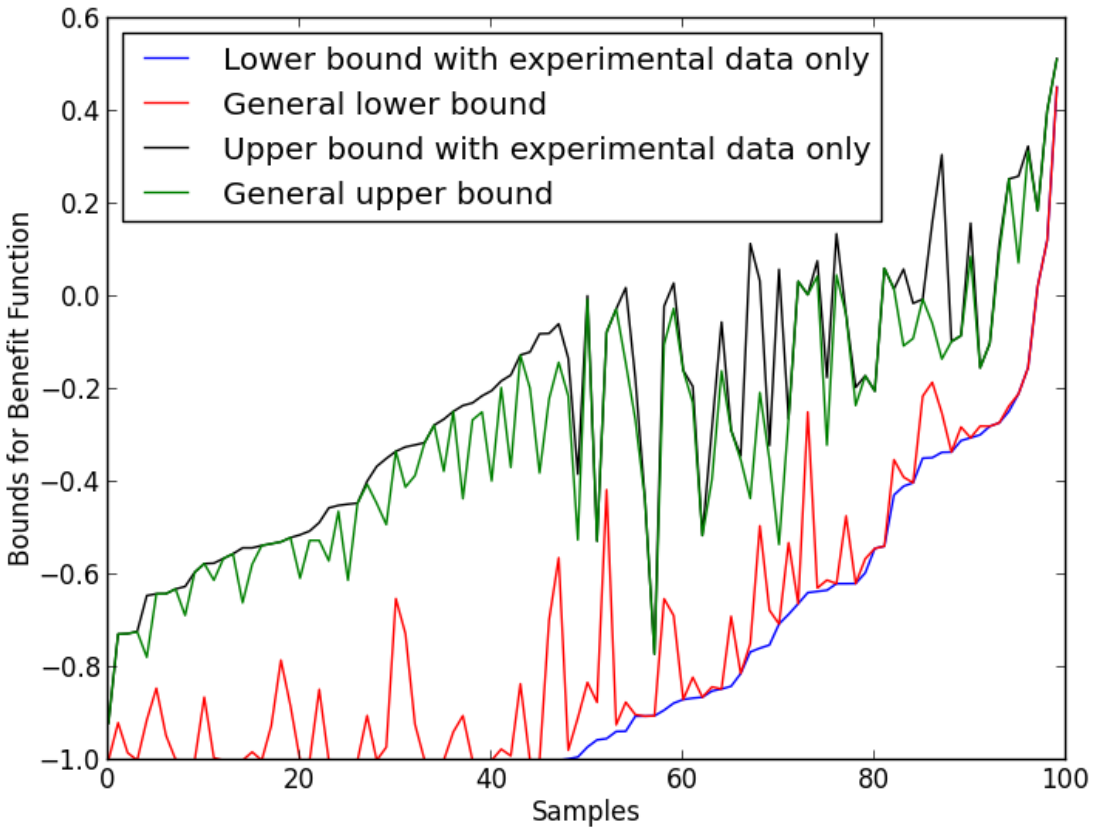

Figure 7.2: Bounds of the benefit function for 100 sample distributions compatible with the causal diagram in Figure 7.1, where the general bounds are obtained from Theorem 4 and the bounds with the experimental data only are obtained from Theorem 12.

Table 7.1: Simulation results of 100000 sample distributions compatible with the causal diagram in Figure 7.1.

Average decreased lower bound	Average increased upper bound	Average gap by Theorem 12	Average gap by Theorem 4	Decision flipped	Bounds wider
0.0569	0.0575	0.5490	0.4346	1148	74834

Theorem 4 suggest gain profit, while the results from Theorem 12 suggest lose profit, or the reverse) are $1148/100000 \approx 1.1\%$ of the samples, which is acceptable. Also, there are $74834/100000 \approx 74.8\%$ of the samples that the bounds with the experimental data only are wider. Therefore, even the bounds are wider in the absence of observational data, the bounds are still valid as they are not affected significantly.

7.2 Unit Selection with Observational Data

In contrast to the situation that only experimental data are available, if there are only observational data, the effectiveness of the bounds are unacceptable if all the experimental data terms in Theorem 4 are removed. Therefore, we need a way to estimate experimental data from observational data. The rest of this chapter investigates how experimental data can be estimated from observational data.

7.2.1 Identifiable Experimental Data

In the case when a causal diagram G is available, and there is either a set of variables that satisfies the back-door criterion in Definition 2 or there is a set of variables that satisfies the front-door criterion in Definition 3, we could estimate experimental data easily using the

adjustment formula in Equation 3.1 or 3.2.

7.2.2 Unidentifiable Experimental Data

In the case that a set of variables that satisfies back-door or front-door criterion is unavailable, we could bound experimental data. The naivest bounds (Tian-Pearl bounds) of experimental data are given in Equation 3.3. In the rest of the chapter, we will discuss that when partially observable back-door or front-front variables are available, we could narrow the bounds of experimental data. The bounds of experimental data would keep getting narrower if we have more back-door or front-door variables being observed, and when the full set of back-door or front-door variables are observed, the bounds shrink into a point estimate. We will demonstrate how the bounds of causal effects (i.e., experimental data, $P(y|do(x))$) can be obtained by non-linear optimizations with partially observable back-door or front-door variables.

7.2.2.1 Partially Observable Back-Door Variables

Theorem 13. *Given a causal diagram G and a distribution compatible with G, let $W \cup U$ be a set of variables satisfying the back-door criterion in G relative to an ordered pair (X, Y), where $W \cup U$ is partially observable, i.e., only probabilities $P(X, Y, W)$ and $P(U)$ are given, the causal effects of X on Y are then bounded as follows:*

$$LB \leq P(y|do(x)) \leq UB$$

where LB is the solution to the non-linear optimization problem in Equation 7.1 and UB is the solution to the non-linear optimization problem in Equation 7.2.

$$LB = \min \sum_{w,u} \frac{a_{w,u} b_{w,u}}{c_{w,u}}, \tag{7.1}$$

$$UB = \max \sum_{w,u} \frac{a_{w,u} b_{w,u}}{c_{w,u}}, \tag{7.2}$$

where,

$\sum_u a_{w,u} = P(x, y, w), \sum_u b_{w,u} = P(w), \sum_u c_{w,u} = P(x, w)$ *for all* $w \in W$;
and for all $w \in W$ *and* $u \in U$,

$b_{w,u} \geq c_{w,u} \geq a_{w,u}$,

$\max\{0, p(x, y, w) + p(u) - 1\} \leq a_{w,u} \leq \min\{P(x, y, w), p(u)\}$,

$\max\{0, p(w) + p(u) - 1\} \leq b_{w,u} \leq \min\{P(w), p(u)\}$,

$\max\{0, p(x, w) + p(u) - 1\} \leq c_{w,u} \leq \min\{P(x, w), p(u)\}$.

7.2.2.2 Partially Observable Front-Door Variables

Theorem 14. *Given a causal diagram G and distribution compatible with G, let* $W \cup U$ *be a set of variables satisfying the front-door criterion in G relative to an ordered pair* (X, Y)*, where* $W \cup U$ *is partially observable, i.e., only probabilities* $P(X, Y, W)$ *and* $P(U)$ *are given and* $P(x, W, U) > 0$*, the causal effects of* X *on* Y *are then bounded as follows:*

$$LB \leq P(y|do(x)) \leq UB$$

where LB is the solution to the non-linear optimization problem in Equation 7.3 and UB is the solution to the non-linear optimization problem in Equation 7.4.

$$LB = \min \sum_{w,u} \frac{b_{x,w,u}}{P(x)} \sum_{x'} \frac{a_{x',w,u} P(x')}{b_{x',w,u}}, \tag{7.3}$$

$$UB = \max \sum_{w,u} \frac{b_{x,w,u}}{P(x)} \sum_{x'} \frac{a_{x',w,u} P(x')}{b_{x',w,u}}, \tag{7.4}$$

where,

$$\sum_{u} a_{x,w,u} = P(x,y,w), \sum_{u} b_{x,w,u} = P(x,w) \text{ for all } x \in X \text{ and } w \in W;$$

and for all $x \in X$,$w \in W$, *and* $u \in U$,

$$b_{x,w,u} \geq a_{x,w,u},$$

$$\max\{0, p(x,y,w) + p(u) - 1\} \leq a_{x,w,u} \leq \min\{P(x,y,w), p(u)\},$$

$$\max\{0, p(x,w) + p(u) - 1\} \leq b_{x,w,u} \leq \min\{P(x,w), p(u)\}.$$

Notably, if any observational data (e.g., $P(U)$) are unavailable in the above theorems, we can remove that term, and the rest of non-linear optimization problems still provide valid bounds for the causal effects. In general, midpoints of bounds on causal effects are effective estimates. However, the lower (upper) bounds are also informative, which can be interpreted as the minimal (maximal) causal effects.

7.2.3 Example

Herein, we present a simulated example to demonstrate that the midpoints of the bounds on causal effects given by Theorem 13 are adequate for estimating the causal effects.

7.2.3.1 Causal Effects of a Drug

Drug manufacturers want to know the causal effect of recovery when a drug is taken. Thus, they conduct an observational study. Here, the recovery rates of 700 patients were recorded.

A total of 192 patients chose to take the drug and 508 patients did not. The results of the study are shown in Table 7.2. Blood type (type O or not) is not the only confounder of taking the drug and recovery. Another confounder is age (below the age of 70 or not). The manufacturers have no data associated with age. They only know that 85.43% of people in their region are below the age of 70.

Table 7.2: Results of an observational study considering blood type.

	Drug	No Drug
Blood type O	23 out of 36 recovered (63.9%)	145 out of 225 recovered (64.4%)
Not blood type O	135 out of 156 recovered (86.5%)	152 out of 283 recovered (53.7%)
Overall	158 out of 192 recovered (82.3%)	297 out of 508 recovered (58.5%)

Because both age and blood type are confounders of taking the drug and recovery, and the data associated with age are unobservable, the causal effect is not identifiable.

Let $X = x$ denote the event that a patient took the drug, $X = x'$ denote the event that a patient did not take the drug, $Y = y$ denote the event that a patient recovered, $Y = y'$ denote the event that a patient did not recover, $W = w$ represent a patient with blood type O, $W = w'$ represent a patient without blood type O, $U = u$ represent a patient below the age of 70, and $U = u'$ represent a patient above the age of 70. The causal diagram is shown in Figure 7.3.

An option for the manufacturers could be estimating the causal effect through the Tian-

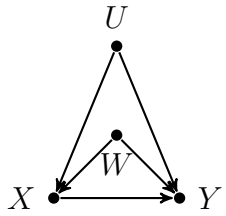

Figure 7.3: Needed the causal effects of X on Y when U is unobserved and independent with W.

Pearl bounds in Equation 3.3 and the observational data from Table 7.2, where

$$\begin{aligned} P(x,y) &= \sum_w P(y|x,w)P(x|w)P(w) = 0.2257, \\ 1-P(x,y') &= 1-\sum_w P(y'|x,w)P(x|w)P(w) = 0.9514. \end{aligned}$$

Therefore, the bounds on the causal effect estimated using Equation 3.3 are $0.2257 \leq P(y|do(x)) \leq 0.9514$, where the causal information of the covariate W and the prior information $P(U)$ are not used. These bounds are not sufficiently informative to conclude the actual causal effect. Although one may believe that we can use the midpoint of the bounds (i.e., 0.5886), the gap (i.e., $0.9514-0.2257 = 0.7257$) between the bounds is not small; hence, this point estimate is unconvincing.

Now, considering the proposed bounds in Theorem 13 with the observational data from Table 7.2. $W \cup U$ satisfies the back-door criterion, and $P(X,Y,W)$ and $P(U)$ are available. We have 12 optimal variables in each objective function, because W and U are binary. With the help of the "SLSQP" solver [Kra88] in the scipy package [Sci20], we obtain the bounds on the causal effect, which are $0.4728 \leq P(y|do(x)) \leq 0.9514$. The lower bound actually increased significantly, and reached close to 0.5, which can help make decisions. The midpoint is 0.7121. Our conclusion is then that the causal effect of recovery when taking the drug is 0.7121. We show in the following section that this estimate of the causal effect is extremely close to the actual causal effect.

7.2.3.2 Informer View of the Causal Effect

An informer with access to the fully observed data, as summarized in Table 7.3 (Note that although it can be verified that the data in Table 7.3 are compatible with those in Table 7.2, we will never know these numbers in reality), would easily calculate the causal effect of recovery when taking the drug using the adjustment formula in Equation 3.1 (shown in Equation 7.5). The error of the estimate of the causal effect using Theorem 13 is only $(0.7518 - 0.7121)/0.7518 \approx 5.28\%$.

$$P(y|do(x)) = \sum_{z,u} P(y|x,z,u)P(z,u) = 0.7518. \tag{7.5}$$

Table 7.3: Informer view of the observational data considering blood type and age.

	Drug	No Drug
Blood type O and Age below 70	3 out of 4 recovered (75.0%)	141 out of 219 recovered (64.4%)
Blood type O and Age above 70	20 out of 32 recovered (62.5%)	4 out of 6 recovered (66.7%)
Not blood type O and Age below 70	135 out of 151 recovered (89.4%)	117 out of 224 recovered (52.2%)
Not blood type O and Age above 70	0 out of 5 recovered (0.0%)	35 out of 59 recovered (59.3%)
Overall	158 out of 192 recovered (82.3%)	297 out of 508 recovered (58.5%)

7.2.4 Simulation Study

Here, we further illustrate that the midpoints of the proposed bounds on the causal effects are sufficient for estimating the causal effects, and the midpoints of the proposed bounds in

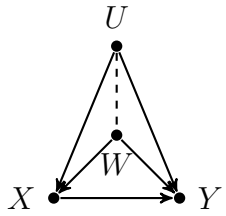

Figure 7.4: Needed the causal effects of X on Y when U is unobserved.

Theorem 13 are better than the midpoints of the Tian-Pearl bounds in Equation 3.3 based on a random simulation.

We employ the simplest causal diagram in Figure 7.4 with binary W, U, such that $W \cup U$ satisfies the back-door criterion. We randomly generated 1000 sample distributions compatible with the causal diagram using Algorithm 2 with random uniform distribution D. The average gap (upper bound $-$ lower bound) of the Tian-Pearl bounds with 1000 samples is 0.487, and the average gap of the proposed bounds with 1000 samples is 0.383. We then randomly picked 100 out of 1000 sample distributions to draw the graph of the actual causal effects, the midpoints of the Tian-Pearl bounds, and the midpoints of the proposed bounds. The results are shown in Figure 7.5.

Algorithm 2: Generate-cpt()

input : n causal diagram nodes $(X_1, ..., X_n)$

Distribution D

output: n conditional probability tables for $P(X_i|Parents(X_i))$

```
begin
    for i ← 1 to n do
        s ← num-instantiates(X_i)
        p ← num-instantiates(Parents(X_i))
        for k ← 1 to p do
            sum ← 0
            for j ← 1 to s do
                a_j ← sample(D)
                sum ← sum + a_j
            end
            for j ← 1 to s do
                P(x_{i_j}|Parents(X_i)_k) ← a_j/sum
            end
        end
    end
end
```

From Figure 7.5, although both midpoints of the bounds on the causal effects are good estimates of the actual causal effects, the midpoints of the proposed bounds are much closer to the actual causal effects, particularly when the causal effects are close to 0 and 1. The average gap (upper bounds − lower bounds), 0.383, of the proposed bounds with 1000 samples is much smaller than the average gap, 0.487, of the Tian-Pearl bounds with 1000 samples. This means that the midpoints of the proposed bounds are more convincing, because the bounds are narrower.

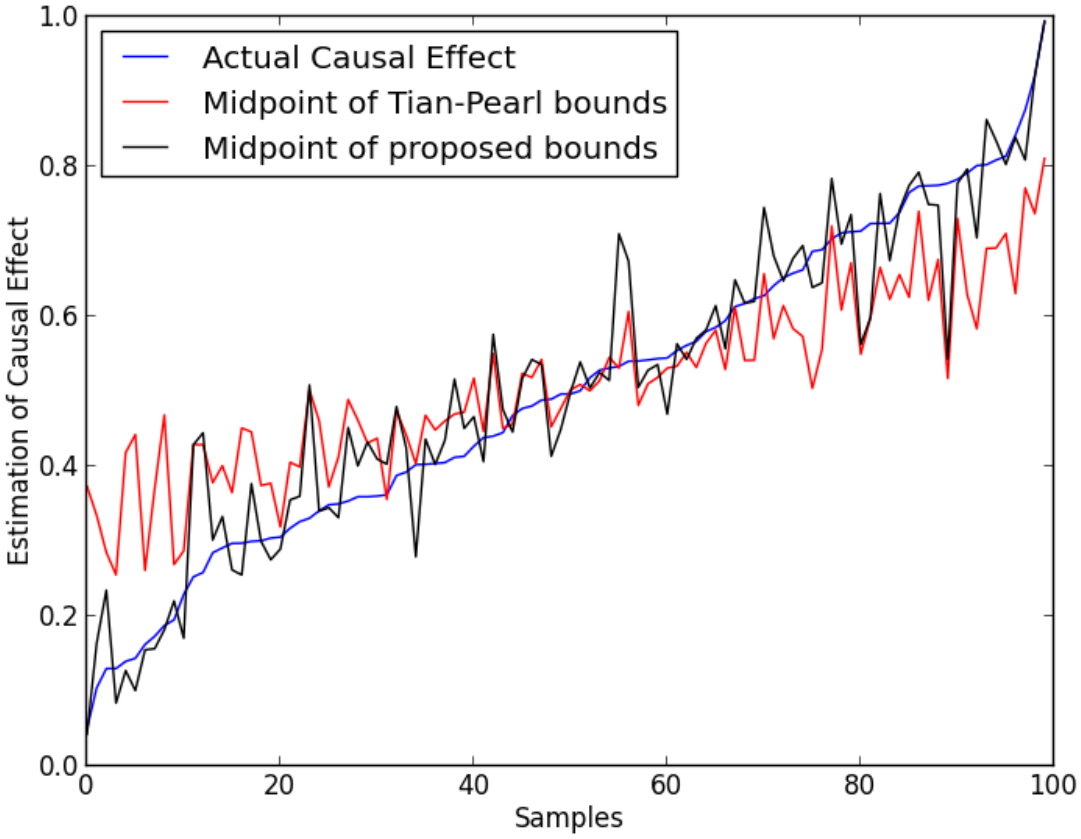

Figure 7.5: Estimates of the causal effects of 100 samples with partially observed confounders, where the Tian-Pearl bounds are obtained from Equation 3.3 and the proposed bounds are obtained through Theorem 13.

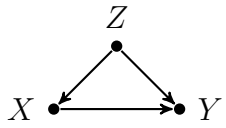

Figure 7.6: Needed the causal effects of X on Y when Z has high dimensionality.

7.2.5 High Dimensionality of Adjustment Variables

Consider the problem of estimating the causal effects of X on Y when a sufficient set Z, which satisfies the back-door or front-door criterion, is fully observable (e.g., see Figure 7.6) in a causal diagram G but has high dimensionality (e.g., Z has 1024 instantiates), a prohibitive large sample size would be required to estimate the causal effects, which is generally recognized to be impractical. Herein, we propose a new framework to achieve dimensionality reduction.

7.2.5.1 Equivalent Causal Diagram with Observational data

Definition 15 (Equivalent causal diagram with observational data)**.** *Let G, G' be causal diagrams both containing nodes X, Y. O are observational data compatible with G, and O' are observational data compatible with G'. We say that (G, O) is equivalent to (G', O') if the causal effects of X on Y with (G, O) is equal to the causal effects of X on Y with (G', O').*

This equivalent tuple (G', O') is easy to obtain. We can simply add two new nodes W and U, and remove a node Z in G to obtain G'. Let the arrows entering Z in G now enter both W and U in G', and let the arrows exiting Z in G now exit both W and U in G'. Finally, add an arrow from U to W. It is easy to show that (G, O) and (G', O') are equivalent if the states of Z are the Cartesian product of the states of W and the states of U. Formally, we have the following theorem,

Theorem 16. *Let G be a causal diagram containing nodes $\{V_1, ..., V_{n-3}, X, Y, Z\}$. Let O be any observational data compatible with G. Suppose there exists a set of variables that satisfies*

the back-door or front-door criterion relative to (X, Y) *in* G*, then,* (G, O) *is equivalent to* (G', O') *(*G' *containing nodes* $\{V_1, ..., V_{n-3}, X, Y, W, U\}$*;* O' *are observational data compatible with* G'*), where the number of states in* W *times the number of states in* U *is equal to the number of states in* Z*, and the structure of* G' *and the observational data* O' *are obtained as follows:*

Structure of G'*:*

Let $Parents_G(H)$ *be the parents of* H *in causal diagram* G*.*

$Parents_{G'}(U) = Parents_G(Z)$,

$Parents_{G'}(W) = Parents_G(Z) \cup \{U\}$.

For $H \in \{V_1, ..., V_{n-3}, X, Y\}$,

$Parents_{G'}(H) = Parents_G(H)$ *if* $Z \notin Parents_G(H)$,

$Parents_{G'}(H) = Parents_G(H) \setminus \{Z\} \cup \{W, U\}$ *if* $Z \in Parents_G(H)$.

Note that, let Q *be the set of variables in* G *that satisfies the back-door or front-door criterion relative to* (X, Y)*, then* Q' *satisfies the back-door or front-door criterion relative to* (X, Y) *in* G' *, where*

$Q' = Q$ *if* $Z \notin Q$,

$Q' = Q \setminus \{Z\} \cup \{W, U\}$ *if* $Z \in Q$.

Observational data:

Let p *be the number of states in* W*, and let* q *be the number of states in* U*.*

The states of Z *are the Cartesian product of the states of* W *and the states of* U*.*

In detail,

(w_j, u_k) *is equivalent to* $z_{(j-1)*q+k}$,

w_j *is equivalent to* $\vee_{k=1}^{q}(w_j, u_k) = \vee_{k=1}^{q} z_{(j-1)*q+k}$,

u_k *is equivalent to* $\vee_{j=1}^{p}(w_j, u_k) = \vee_{j=1}^{p} z_{(j-1)*q+k}$,

$P(w_j, u_k, V) = P(z_{(j-1)*q+k}, V)$ *for any* $V \subseteq \{V_1, ..., V_{n-3}, X, Y\}$.

For example, consider the causal diagram in Figure 7.6 and the observational data in

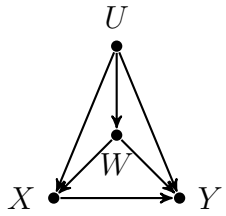

Figure 7.7: Causal diagram of an equivalent problem.

Table 7.5 (in the form of conditional probability tables (CPTs), where X, Y are binary, and Z has 4 states). The causal effect, $P(y|do(x))$, through the adjustment formula in Equation 3.1, is 0.47. Based on the construction (details are shown in Table 7.4) in Theorem 16, we have the causal diagram in Figure 7.6 with the observational data in Table 7.5 is equivalent to the causal diagram in Figure 7.7 with the observational data in Table 7.6 (all nodes are binary), and we can verify that the causal effect, $P(y|do(x))$, in the causal diagram in Figure 7.7 with the observational data in Table 7.6 is also 0.47.

Notably, the equivalent tuple is not unique and is transitive (i.e., if (G, O) is equivalent to (G', O'), and (G', O') is equivalent to (G'', O''), then (G, O) is equivalent to (G'', O'')).

7.2.5.2 Dimensionality Reduction

Now, consider the problem in the beginning of Section 7.2.5. First, we transform the causal diagram G with compatible observational data O into an equivalent tuple (G', O') using Algorithm 3 based on the construction in Theorem 16 (note that the algorithm only construct the structure of G' and assigning meanings to the states W and U, the corresponding observational data O' are then easy to obtain), then the new problem (G', O') has the same causal effects of X on Y as in (G, O). By picking the dimensionality of W (p in Algorithm 3), we can control the dimensionality of the new problem.

Note that, if $Z = (Z_1, Z_2, ..., Z_m)$ in G is a set of variables, we can repeat Algorithm 3 for each variable in Z, and finally obtain $W = (W_1, W_2, ..., W_m)$ and $U = (U_1, U_2, ..., U_m)$,

Table 7.4: Construction of the observational data based on Theorem 16.

$P(u, w) = P(z_1)$
$P(u, w') = P(z_2)$
$P(u', w) = P(z_3)$
$P(u', w') = P(z_4)$
$P(u) = P(u, w) + P(u, w') = P(z_1) + P(z_2) = 0.5$
$P(w\|u) = P(u, w)/p(u) = P(z_1)/P(u) = 0.3/0.5 = 0.6$
$P(w\|u') = P(u', w)/p(u') = P(z_3)/(1 - P(u)) = 0.2/0.5 = 0.4$
$P(x\|u, w) = P(x\|z_1) = 0.1$
$P(x\|u, w') = P(x\|z_2) = 0.4$
$P(x\|u', w) = P(x\|z_3) = 0.5$
$P(x\|u', w') = P(x\|z_4) = 0.7$
$P(y\|x, u, w) = P(y\|x, z_1) = 0.2$
$P(y\|x', u, w) = P(y\|x', z_1) = 0.3$
$P(y\|x, u, w') = P(y\|x, z_2) = 0.7$
$P(y\|x', u, w') = P(y\|x', z_2) = 0.1$
$P(y\|x, u', w) = P(y\|x, z_3) = 0.6$
$P(y\|x', u', w) = P(y\|x', z_3) = 0.5$
$P(y\|x, u', w') = P(y\|x, z_4) = 0.5$
$P(y\|x', u', w') = P(y\|x', z_4) = 0.4$

where the multiplication of the number of states in W is equal to p.

We then treat the new problem (G', O') as a partially observable back-door or front-door variables problem in Sections 7.2.2.1 and 7.2.2.2, where $P(X, Y, W)$ and $P(U)$ are given, and we can then obtain the bounds of the causal effects through Theorems 13 and 14. We claim that the midpoints of the bounds are good estimates of the original causal effects. In addition, the bounds themselves will help make decisions.

Table 7.5: Observational data in CPTs compatible with the causal diagram in Figure 7.6.

$P(z_1)$	0.3
$P(z_2)$	0.2
$P(z_3)$	0.2
$P(z_4)$	0.3

$P(x\|z_1)$	0.1
$P(x\|z_2)$	0.4
$P(x\|z_3)$	0.5
$P(x\|z_4)$	0.7

$P(y\|x, z_1)$	0.2
$P(y\|x', z_1)$	0.3
$P(y\|x, z_2)$	0.7
$P(y\|x', z_2)$	0.1
$P(y\|x, z_3)$	0.6
$P(y\|x', z_3)$	0.5
$P(y\|x, z_4)$	0.5
$P(y\|x', z_4)$	0.4

Table 7.6: Observational data in CPTs compatible with the causal diagram in Figure 7.7.

$P(u)$	0.5

$P(w\|u)$	0.6
$P(w\|u')$	0.4

$P(x\|u, w)$	0.1
$P(x\|u, w')$	0.4
$P(x\|u', w)$	0.5
$P(x\|u', w')$	0.7

$P(y\|x, u, w)$	0.2
$P(y\|x', u, w)$	0.3
$P(y\|x, u, w')$	0.7
$P(y\|x', u, w')$	0.1
$P(y\|x, u', w)$	0.6
$P(y\|x', u', w)$	0.5
$P(y\|x, u', w')$	0.5
$P(y\|x', u', w')$	0.4

7.2.5.3 Example

Considering the problem in Figure 7.6, where X and Y are binary and Z has 256 states. We randomly generated a distribution $P(X, Y, Z)$ that is compatible with the causal diagram using Algorithm 2. Because we know the exact distribution, we can easily obtain the causal effects through Equation 3.1. The causal effect $P(y|do(x))$ is 0.5527.

Algorithm 3: Generate Equivalent Tuple

input : A n nodes, $(X_1, X_2, ..., X_{n-3}, X, Y, Z)$, causal diagram G and compatible O, p, the number of states in W in G' of the equiv. tuple (G', O').

output: A $n+1$ nodes, $(X_1, X_2, ..., X_{n-3}, X, Y, W, U)$, causal diagram G', Maping relation M_1 : state of W $\rightarrow$ state of Z, Maping relation M_2 : state of U $\rightarrow$ state of Z.

begin

$m \leftarrow$ `num_states_in_G(`Z`);`

if $m \mod p = 0$ **then**

$q \leftarrow m/p$;

end

else

$q \leftarrow m/p + 1$;

end

/ / Set the virtual states for Z such that the probaility is 0;

`num_states_in_G(`Z`)` $\leftarrow p \times q$;

for H *in* $\{X_1, ..., X_{n-3}, X, Y\}$ **do**

`num_states_in_G'(`H`)` $\leftarrow$ `num_states_in_G(`H`);`

if $Z \in$`Parents_in_G(`H`)` **then**

`Parents_in_G'(`H`)` $\leftarrow$ `Parents_in_G(`H`)` $\setminus \{Z\} \cup \{W, U\}$;

end

else

`Parents_in_G'(`H`)` $\leftarrow$ `Parents_in_G(`H`);`

end

end

`num_states_in_G'(`W`)` $\leftarrow p$;

`num_states_in_G'(`U`)` $\leftarrow q$;

`Parents_in_G'(`W`)` $\leftarrow$ `Parents_in_G(`Z`)` $\cup \{U\}$;

`Parents_in_G'(`U`)` $\leftarrow$ `Parents_in_G(`Z`);`

for $i \leftarrow 1$ **to** p **do**

$M_1(w_i) \leftarrow \vee_{k=1}^{q} z_{(i-1)*q+k}$;

end

for $i \leftarrow 1$ **to** q **do**

$M_2(u_i) \leftarrow \vee_{j=1}^{p} z_{(j-1)*q+i}$;

end

end

Now, we transform the causal diagram with the observational data into an equivalent tuple (G', O') (G' is shown in Figure 7.7) using Algorithm 3 ($p = 16$). We obtain a variable W of 16 states and a variable U of 16 states in G' ((w_j, u_k) is equivalent to $z_{(j-1)*16+k}$). We are then forced to use only observational data $P(X, Y, W)$ and $P(U)$ (the construction of the data is shown in the next paragraph), and based on Theorem 13, with the "SLSQP" solver, we obtain the bounds on the causal effect, which are $0.4595 \leq P(y|do(x)) \leq 0.7012$. We see the midpoint, 0.5804, is extremely close to the actual causal effect, 0.5527.

Instead of providing the resulting 1024 rows of the observational data, we provide the details for regenerating the observational data as following steps:

- Generate $P(X, Y, Z)$ using Algorithm 2;
- Let $P(X, Y, w_j, u_k) = P(X, Y, z_{(j-1)*16+k})$;
- Let $P(X, Y, w_j) = \sum_{k=1}^{q} P(X, Y, w_j, u_k)$;
- Let $P(X, Y, u_k) = \sum_{j=1}^{p} P(X, Y, w_j, u_k)$;
- Let $P(u_k) = \sum_{X,Y} P(X, Y, u_k)$.

For example,

$$
\begin{aligned}
P(u_1) &= \sum_{X,Y} P(X, Y, u_1) \\
&= P(x, y, u_1) + P(x, y', u_1) + P(x', y, u_1) + P(x', y', u_1) \\
&= \sum_{j=1}^{16} P(x, y, w_j, u_1) + \sum_{j=1}^{16} P(x, y', w_j, u_1) + \\
&\quad + \sum_{j=1}^{16} P(x', y, w_j, u_1) + \sum_{j=1}^{16} P(x', y', w_j, u_1) \\
&= \sum_{j=1}^{16} P(x, y, z_{(j-1)*16+1}) + \sum_{j=1}^{16} P(x, y', z_{(j-1)*16+1}) + \\
&\quad \sum_{j=1}^{16} P(x', y, z_{(j-1)*16+1}) + \sum_{j=1}^{16} P(x', y', z_{(j-1)*16+1}), \\
P(x, y, w_1) &= \sum_{k=1}^{16} P(x, y, w_1, u_k) \\
&= \sum_{k=1}^{16} P(x, y, z_k).
\end{aligned}
$$

Finally, lets consider how many samples are required for each method. According to [Ros75], each state needs at least 30 samples, and therefore, the exact solution by Equation 3.1 requires $2 \times 2 \times 256 \times 30 = 30720$ samples. However, the proposed bounds based on Theorem 13 only requires $\max(2 \times 2 \times 16, 16) \times 30 = 1920$ samples. If the sample size is still unacceptable, we can use another equivalent tuple with W having 8 states and U having 32 states, we then only require $\max(2 \times 2 \times 8, 32) \times 30 = 960$ samples to obtain the bounds on the causal effects.

7.2.5.4 Simulation Study

Similarly to the previous simulation, we further illustrate that the bounds on the causal effects of the proposed framework are sufficient for estimating the original causal effects.

Once again, by employing the simplest causal diagram in Figure 7.6, where X and Y are

binary and Z has 256 states. We randomly generated 100 sample distributions compatible with the causal diagram using Algorithm 2. The average gap (upper bound − lower bound) of the Tian-Pearl bounds in Equation 3.3 with 100 samples is 0.5102, and the average gap of the proposed bounds through Theorems 16 and 13 with 100 samples is 0.0676. We then draw the graph of the actual causal effects, the midpoints of the Tian-Pearl bounds, and the midpoints of the proposed bounds through Theorems 16 and 13. The results are shown in Figure 7.8.

From Figure 7.8, both midpoints of the bounds on the causal effects are good estimates of the actual causal effects, whereas the midpoints of the proposed bounds are slightly closer to the actual causal effects, particularly when the causal effects are close to 0 and 1. Although the trend of the Tian-Pearl bounds is also close to the actual causal effects, the Tian-Pearl bounds are more likely to be parallel with the x-axis. Here, the Tian-Pearl bounds perform well because, in high-dimensionality cases, the randomly generated distributions are more likely to yield causal effects of approximately 0.5. However, the average gap of the proposed bounds with 100 samples, 0.0676, is much smaller than the average gap of the Tian-Pearl bounds with 100 samples, 0.5102. This means that the midpoints of the proposed bounds are more convincing, because the bounds are narrower.

7.2.6 Discussion

Here, we discuss additional features of bounds on causal effects.

First, if a whole set of back-door or front-door variables are unobserved, the causal effects have the naivest bounds in Equation 3.3. When the back-door or front-door variables are gradually observed, the bounds of the causal effects become increasingly narrow. Finally, when the back-door or front-door variables are fully observed, the bounds shrink into point estimates, which are identifiable. This also tells us that, when we pick p in Algorithm 3, we should pick the largest p for which the sample size is sufficient to estimate the observational distributions.

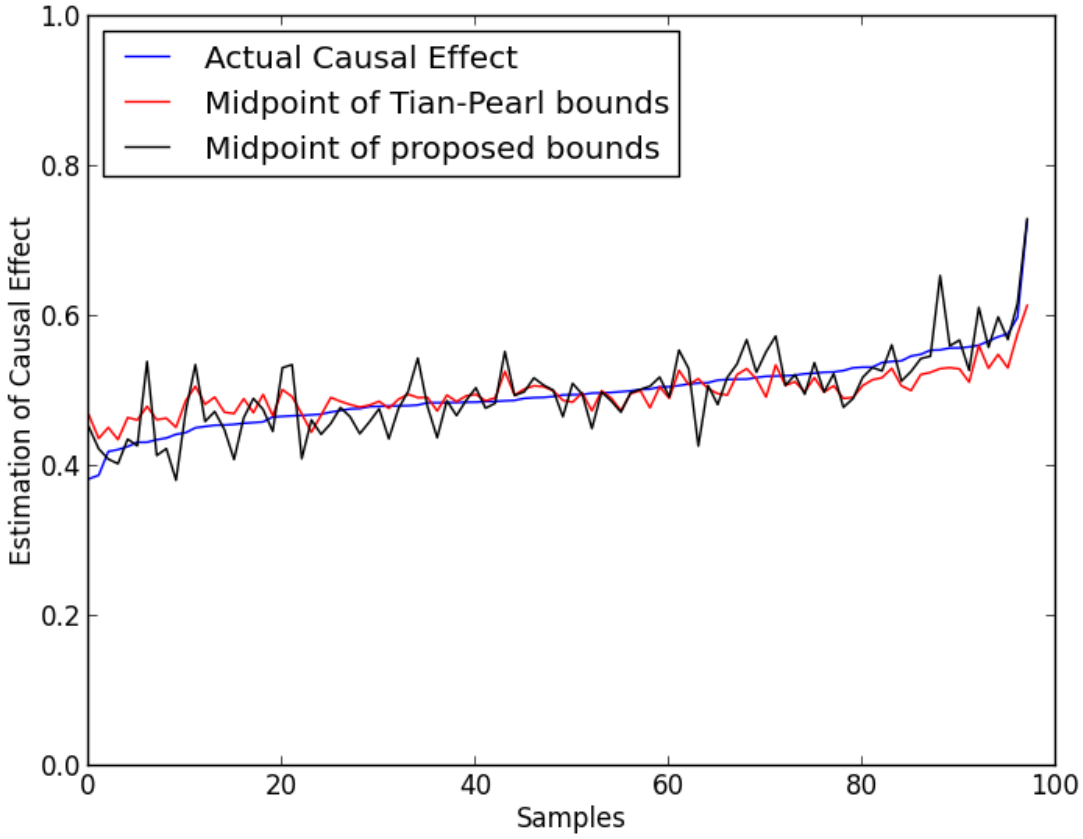

Figure 7.8: Estimates of the causal effects of 100 samples with high dimensionality data, where the Tian-Pearl bounds are obtained from Equation 3.3 and the proposed bounds are obtained through Theorems 16 and 13.

Next, the bounds in Theorems 13 and 14 are given by non-linear optimizations. Therefore, the quality of the bounds also depends on the optimization solver. The examples and simulated results in this paper are all obtained from the simplest "SLSQP" solver from 1988. The quality of the bounds can be improved if more advanced solvers are applied. Inspired by the idea of Balke's linear programming [BP97b], we may obtain parametric solutions to non-linear optimizations in Theorems 13 and 14, we then do not need a non-linear optimization solver. However, the problem related to a non-linear optimization solver is not the scope of this study.

In addition, the constraints in Theorems 13 and 14 are only based on the basic back-door or front-door criterion. We can also add constraints of independencies in a specific graph. For instance, W and U are independent in the causal diagram of Figure 7.3, we can then add the constraints that reflect $P(W)$ and $P(U)$ as being independent. The greater the number of constraints that are added to the optimizations, the better the bounds we can obtain.

Moreover, if one believes they have a sufficient sample size to estimate causal effects with high dimensionality adjustment variables, the framework in Section 7.2.5 could be evidence validating whether the sample size is indeed sufficient.

Next, in Section 7.2.5, we transformed (G, O) into (G', O') to obtain bounds on causal effects with high dimensionality adjustment variables. However, for a tuple (G, O), multiple equivalent tuples exist by picking the different p in Algorithm 3, and each of the equivalent tuple has bounds for the original causal effects. We can compute the bounds for as many equivalent tuples as we want and take the maximal lower bounds and the minimal upper bounds.

Finally, based on numerous experiments, we realized that when $P(U)$ or $P(W)$ is specific (i.e., closer to 0 or 1), the proposed bounds are almost identified (i.e., the bounds shrink to point estimates). Therefore, in practice, we can always pick the equivalent tuple to transform, in which the $P(U)$ or $P(W)$ is close to 0 or 1.

CHAPTER 8

Applications

Recall that the benefit vector in Chapter 4 is not determined by the model but by the one who uses the model. In this chapter, we illustrate several common applications showing how to set the benefit vector. We categorize the applications based on the quality of A/B-test-based approaches.

8.1 Cases in which Simple A/B-test-based Approaches are Correct

8.1.1 Number of Increased Customers

Consider a mobile carrier that wants to identify customers likely to discontinue their services within the next quarter based on customer characteristics (the company management has access to user data, such as income, age, usage, and monthly payments). The carrier will then offer these customers a special renewal deal to dissuade them from discontinuing their services and to increase their service renewal rate.

Let $A = a$ denote the event that a customer receives the special deal, $A = a'$ denote the event that a customer receives no special deal, $R = r$ denote the event that a customer continues the services, $R = r'$ denote the event that a customer discontinues the services, and C (a set of variables) denote the population-specific characteristics of a customer (e.g., income, age, usage, and monthly payments).

If the manager only wants to maximize the number of increased customers due to the offer in the next quarter regardless of the total profit, then they should assign 1 to a complier because the company gains one customer due to the offer, assign 0 to an always-taker and a never-taker because the company gains no customer due to the offer, and assign -1 to a defier because the company loses one customer due to the offer.

Therefore, the benefit vector above is $(1, 0, 0, -1)$, and using Theorem 8, when the benefit vector satisfies the gain equality $(1 - 1 = 0 + 0)$, the benefit function is $f(c) = P(r_a|c) - P(r_{a'}|c)$. This is the most common A/B test heuristic in literature.

8.1.2 Number of Total Customers

If the manager only wants to maximize the total number of customers in the next quarter regardless of the total profit, then they should assign 1 to a complier and an always-taker because the company has one customer in the next quarter and assign 0 to a never-taker and a defier because the company has no customer in the next quarter.

Therefore, the benefit vector above is $(1, 1, 0, 0)$, and using Theorem 8, when the benefit vector satisfies the gain equality $(1 + 0 = 1 + 0)$, the benefit function is $f(c) = P(r_a|c)$. This is another common A/B test heuristic in literature, which is the causal effect of the offer to the number of customers.

8.1.3 Immediate Profit

If the manager wants to maximize the total immediate profit due to the offer. The management estimates that the benefit of selecting a complier is \$100 as the profit is \$140 but the discount is \$40, the benefit of selecting an always-taker is $-$\$40 as the customer would continue the service anyway and the company loses the value of the discount, the benefit of selecting a never-taker is \$0 as the cost of issuing the discount is negligible, and the benefit of selecting a defier is $-$\$140 as they lose a customer due to the special offer.

Therefore, the benefit vector above is $(100, -40, 0, -140)$, using Theorem 8, when the benefit vector satisfies the gain equality $(100 - 140 = -40 + 0)$, the benefit function is $f(c) = 100P(r_a|c) - 140P(r_{a'}|c)$. This result is the same as the popular method in the industry, which is called revenue difference. The profit of a continuing customer if issued the special offer is \$100 and the profit of a continuing customer if no special offer is issued is \$140; therefore, the revenue difference is $100P(r_a|c) - 140P(r_{a'}|c)$.

8.2 Cases in which Simple A/B-test-based Approaches are not Correct

8.2.1 Nonimmediate Profit

If the manager wants to maximize the total profit including the nonimmediate profit due to the offer. The management estimates that the benefit of selecting a complier is \$100 as the profit is \$140 but the discount is \$40, the benefit of selecting an always-taker is −\$60 as the customer would continue the service anyway (so the company loses the value of the discount and an extra cost \$20 because the always-taker may require additional discounts in the future), the benefit of selecting a never-taker is 0 as the cost of issuing the discount is negligible, and the benefit of selecting a defier is −\$140 as they lose a customer due to the special offer.

Therefore, the benefit vector above is $(100, -60, 0, -140)$, and this is the example we have illustrated in Chapter 5.3.1, where the simple A/B-test-based approach is NOT correct.

8.2.2 Minimize the Number of Ineffective Patients and the Number of Serious Side-effect Patients

A pharmaceutical factory invents a new medicine and wants to identify patients so as to minimize the number of ineffective patients plus the number of patients who have serious

side-effect, with focus on the patients who have serious side-effect (the side-effect may lead to death).

Therefore, they should assign unit 0 to a complier because the complier is the patient cured by the medicine, assign unit -1 to an always-taker and a never-taker because the always-taker and never-taker are the ineffective patients that do not respond to the medicine, and assign unit -2 to a defier because the defier is the patient who have serious side-effect (may lead to death).

Let $A = a$ denote the event that a patient receives the medicine, $A = a'$ denote the event that a patient receives no medicine, $R = r$ denote the event that a patient recovered, $R = r'$ denote the event that a patient does not recover, and C (a set of variables) denote the population-specific characteristics of a patient.

The benefit vector above is $(0, -1, -1, -2)$, using Theorem 8, when the benefit vector satisfies the gain equality $(0-2 = -1-1)$, the benefit function is $f(c) = P(r_a|c) - P(r_{a'}|c) - 1 = -P(r'_a|c) - P(r_{a'}|c)$.

Notably, even the benefit function is a point estimate and only requires experimental data but it is difficult to determine the coefficients using a simple A/B-test-based approach.

8.2.3 Maximize Users Satisfaction

The management of a search engine company wants to decide whether it is worth sending an advertisement to a group of users, so as to maximize overall satisfaction. The management estimates that the satisfaction of recommending an advertisement to a complier is 2 degrees, as users would gain new information that they needed, that of recommending the advertisement to an always-taker is 1 degree, as users got a shortcut to the advertisement, that of recommending the advertisement to a never-taker is -1 degrees, as users got unnecessary information, and that of recommending the advertisement to a defier is -2 degrees, as the recommendation would prevent users to get needed information (compliers are the

users who would click on the advertisement if the advertisement is recommended and would not if otherwise; always-takers are the users who would click on the advertisement whether or not the advertisement is recommended; never-takers are the users who would not click on the advertisement whether or not the advertisement is recommended; defiers are the users who would click on the advertisement if the advertisement is not recommended and would not if otherwise).

Therefore, the benefit vector above is $(2, 1, -1, -2)$, and this is the example we have illustrated in Chapter 5.3.2, where a simple A/B-test-based approach is NOT correct because the coefficients are difficult to be determined.

8.2.4 Maximize Difference between the Number of Effective Patients and the Number of Ineffective Patients

A pharmaceutical factory invents a new medicine and wants to identify patients so as to maximize difference between the number of effective patients and the number of ineffective patients.

Therefore, they should assign 1 to a complier because the complier is the patient cured by the medicine, assign -1 to an always-taker, a never-taker, and a defier because they are all ineffective patients. The benefit vector is then $(1, -1, -1, -1)$.

Let $A = a$ denote the event that a patient receives the medicine, $A = a'$ denote the event that a patient receives no medicine, $R = r$ denote the event that a patient is cured, $R = r'$ denote the event that a patient is not cured, and C (a set of variables) denote the population-specific characteristics of a patient.

Suppose they have two groups of patients, group 1 with characteristics c_1 and group 2 with characteristics c_2. In addition, they have prior information that $P(r|c_1) = 0.3$ and $P(r|c_2) = 0.1$. They randomly select 700 patients from each group and offer the medicine to 350 customers in each group. Table 8.1 summarizes the results.

Table 8.1: Results of a simulated study on patients.

		$do(a)$	$do(a')$
Group 1	r	210	105
	r'	140	245
Group 2	r	217	129
	r'	133	221

Table 8.2: Results of the two objective functions based on the data from the simulated study.

	f_1	f_2	real
Group 1	0.3	-0.1	-0.2
Group 2	0.25	0.14	0.2

Let us compare the two selection strategies, each using a different objective function. The first is a simple A/B test heuristic, that is:

$$\begin{aligned} Obj_1 &= argmax_c f_1(c) \\ &= argmax_c P(r|c, do(a)) - P(r|c, do(a')). \end{aligned}$$

The second is the proposed approach, that is:

$$\begin{aligned} Obj_2 &= argmax_c f_2(c) \\ &= argmax_c 1 \times P(r_a, r'_{a'}|c) + (-1) \times P(r_a, r_{a'}|c) + \\ &\quad (-1) \times P(r'_a, r'_{a'}|c) + (-1) \times P(r'_a, r_{a'}|c). \end{aligned}$$

Then, we enter the data in Table 8.1 into the objective functions of groups 1 and 2. Table 8.2 summarizes the results (note that we use the midpoint of the bounds from Theorem 4 as the selection criterion for Obj_2). The proposed approach selected group 2; however, the first objective function selected group 1 as the desired patients.

Table 8.3: Percentages of four response types in each group for patients.

	Complier	Always-taker	Never-taker	Defier
Group 1	40%	20%	30%	10%
Group 2	60%	2%	3%	35%

An informer with access to the fractions of compliers, always-takers, never-takers, and defiers in both groups (as summarized in Table 8.3, and these numbers are never known in reality) would easily conclude that the A/B test heuristic had reached a wrong conclusion. In detail, the expected benefit of selecting a patient in group 1 is $1\times0.4-1\times0.2-1\times0.3-1\times0.1 = -0.2$, which means offering the medicine to group 1 would have negative difference; the expected benefit of selecting a patient in group 2 is $1\times0.6-1\times0.02-1\times0.03-1\times0.35 = 0.2$. Thus, the pharmaceutical factory should only offer the medicine to group 2.

CHAPTER 9

Conclusion

We demonstrated the advantages of the SCM framework in addressing the unit selection problem. We defined an objective function for selection that properly accounts for the counterfactual nature of the desired behavior. We derived tight bounds (Theorem 4) to ensure that the objective function can be evaluated using experimental and observational data. We further identified using Theorem 8 the conditions under which the standard A/B test heuristic used in the literature can become optimal. We then provided the graphical conditions (Theorem 9, 10, and 11) such that the bounds of the objective function can be narrower with additional covariates. We also discussed data availability issue, i.e., how to evaluate the objective function in the absence of either observational data or experimental data. We finally demonstrated how to set up the benefit vector by applications. In summary, we have analyzed and demonstrated what can be gained by exploiting causal knowledge, when solving the unit selection problem.

APPENDIX A

Proofs for Chapter 5

Lemma 17. *The c-specific PNS $P(y_x, y'_{x'}|c)$ is bounded as follows:*

$$max\left\{\begin{array}{c} 0, \\ P(y_x|c) - P(y_{x'}|c), \\ P(y|c) - P(y_{x'}|c), \\ P(y_x|c) - P(y|c) \end{array}\right\} \leq c\text{-}PNS, \tag{A.1}$$

$$min\left\{\begin{array}{c} P(y_x|c), \\ P(y'_{x'}|c), \\ P(y, x|c) + P(y', x'|c), \\ P(y_x|c) - P(y_{x'}|c) + P(y, x'|c) + P(y', x|c) \end{array}\right\} \geq c\text{-}PNS. \tag{A.2}$$

Proof. Since for any three events A, B and C, we have,

$$P(A, B|C) \geq max[0, P(A|C) + P(B|C) - 1], \tag{A.3}$$

therefore, we have,

$$\begin{aligned} c\text{-PNS} &\geq max[0, P(y_x|c) + P(y'_{x'}|c) - 1] \\ &= max[0, P(y_x|c) - P(y_{x'}|c)]. \end{aligned}$$

Also,

$$\begin{aligned}
c\text{-PNS} &= P(y_x, y'_{x'}, x|c) + P(y_x, y'_{x'}, x'|c) \\
&= P(y, y'_{x'}, x|c) + P(y_x, y', x'|c) \qquad \text{(A.4)} \\
&= P(x, y|c) - P(x, y, y_{x'}|c) + P(y_x, y', x'|c) \\
&= P(x, y|c) - P(y, y_{x'}|c) + P(x', y, y_{x'}|c) + P(y_x, y', x'|c) \\
&= P(x, y|c) - P(y, y_{x'}|c) + P(x', y|c) + P(y_x, y', x'|c) \\
&= P(y|c) - P(y, y_{x'}|c) + P(x', y', y_x|c) \qquad \text{(A.5)} \\
&= P(y|c) - P(y, y_{x'}|c) + P(y', y_x|c) - P(x, y', y_x|c) \\
&= P(y|c) - P(y, y_{x'}|c) + P(y', y_x|c) - P(x, y', y|c) \\
&= P(y|c) - P(y, y_{x'}|c) + P(y', y_x|c). \qquad \text{(A.6)}
\end{aligned}$$

By (A.6),

$$\begin{aligned}
c\text{-PNS} &\geq P(y|c) - P(y, y_{x'}|c) \\
&\geq P(y|c) - P(y_{x'}|c).
\end{aligned}$$

Also by (A.6) and (A.3),

$$\begin{aligned}
c\text{-PNS} &\geq P(y|c) - P(y|c) + P(y', y_x|c) \\
&\geq P(y'|c) - P(y'_x|c) \\
&= P(y_x|c) - P(y|c).
\end{aligned}$$

Thus, the lower bounds are proved.

And since for any three events A, B and C, we have,

$$P(A, B|C) \leq min[P(A|C), P(B|C)], \qquad \text{(A.7)}$$

therefore, we have,

$$c\text{-PNS} \leq min[P(y_x|c), P(y'_{x'}|c)].$$

Also, by (A.4),

$$\text{c-PNS} \leq P(x, y|c) + P(x', y'|c).$$

Similarly to (A.5), we have,

$$\begin{aligned}
c\text{-PNS} &= P(y'|c) - P(y', y'_x|c) + P(x, y, y'_{x'}|c) \\
&= P(y', y_x|c) + P(x, y, y'_{x'}|c) \\
&= P(y_x|c) - P(y, y_x|c) + P(x, y, y'_{x'}|c) \\
&= P(y_x|c) - P(y, y_x|c) + P(x, y|c) - P(x, y, y_{x'}|c) \\
&= P(y_x|c) - P(y, y_x|c) + P(x, y|c) - P(y_{x'}|c) + P(x', y, y_{x'}|c) + \\
&\quad P(x, y', y_{x'}|c) + P(x', y', y_{x'}|c) \\
&= P(y_x|c) - P(y, y_x|c) + P(x, y|c) - P(y_{x'}|c) + P(x', y|c) + P(x, y', y_{x'}|c) \\
&= P(y_x|c) - P(y_{x'}|c) + P(x', y|c) + P(x, y|c) - P(y, y_x|c) + P(x, y', y_{x'}|c) \\
&= P(y_x|c) - P(y_{x'}|c) + P(x', y|c) + P(x, y|c) - P(x, y, y_x|c) - P(x', y, y_x|c) + \\
&\quad P(x, y'|c) - P(x, y', y'_{x'}|c) \\
&= P(y_x|c) - P(y_{x'}|c) + P(x', y|c) + P(x, y'|c) - P(x, y', y'_{x'}|c) - P(x', y, y_x|c) \\
&\leq P(y_x|c) - P(y_{x'}|c) + P(x', y|c) + P(x, y'|c).
\end{aligned}$$

Thus, the upper bounds are proved. □

Lemma 18.

$$\begin{aligned}
&P(y_x, y'_{x'}|c) - P(y'_x, y_{x'}|c) \\
= \;& P(y_x|c) - P(y_{x'}|c) \qquad \text{(A.8)}
\end{aligned}$$

Proof.

$$
\begin{aligned}
& P(y_x, y'_{x'}|c) - P(y_{x'}, y'_x|c) \\
= \;& P(y_x, y'_{x'}, x|c) + P(y_x, y'_{x'}, x'|c) - P(y_{x'}, y'_x, x|c) - P(y_{x'}, y'_x, x'|c) \\
= \;& P(y, y'_{x'}, x|c) + P(y_x, y', x'|c) - P(y_{x'}, y', x|c) - P(y, y'_x, x'|c) \\
= \;& P(y, y'_{x'}, x|c) - P(y_{x'}, y', x|c) + P(y_x, y', x'|c) - P(y, y'_x, x'|c) \\
= \;& P(x, y|c) - P(y, y_{x'}, x|c) - P(y_{x'}, y', x|c) + P(y_x, y', x'|c) + P(y, y_x, x'|c) - P(x', y|c) \\
= \;& P(x, y|c) - P(y_{x'}, x|c) + P(y_x, x'|c) - P(x', y|c) \\
= \;& P(x, y|c) - P(y_{x'}|c) + P(y_{x'}, x'|c) + P(y_x|c) - P(y_x, x|c) - P(x', y|c) \\
= \;& P(x, y|c) - P(y_{x'}|c) + P(y, x'|c) + P(y_x|c) - P(y, x|c) - P(x', y|c) \\
= \;& P(y_x|c) - P(y_{x'}|c).
\end{aligned}
$$

□

Lemma 19. *The counterfactual expression* $f(\alpha) = \alpha P(y_x, y'_{x'}|c) - (1-\alpha)P(y_{x'}, y'_x|c)$ *for any real number* α *is bounded as follows.*

Case 1: $\alpha \in (-\infty, 0.5)$,

$$
max \left\{ \begin{array}{c}
\alpha P(y_x|c) - (1-\alpha)P(y_{x'}|c), \\
(1-\alpha)P(y_x|c) + \alpha P(y'_{x'}|c) + \alpha - 1, \\
(2\alpha-1)P(y, x|c) + (2\alpha-1)P(y', x'|c)] + (1-\alpha)[P(y_x|c) - P(y_{x'}|c)], \\
\alpha[P(y_x|c) - P(y_{x'}|c)] + (2\alpha-1)P(y, x'|c) + (2\alpha-1)P(y', x|c)
\end{array} \right\} \leq f(\alpha), \tag{A.9}
$$

$$
min \left\{ \begin{array}{c}
(1-\alpha)[P(y_x|c) - P(y_{x'}|c)], \\
\alpha[P(y_x|c) - P(y_{x'}|c)], \\
(2\alpha-1)P(y|c) + (1-\alpha)P(y_x|c) - \alpha P(y_{x'}|c), \\
\alpha P(y_x|c) - (1-\alpha)P(y_{x'}|c) - (2\alpha-1)P(y|c)
\end{array} \right\} \geq f(\alpha). \tag{A.10}
$$

Case 2: $\alpha \in [0.5, \infty)$,

$$max \left\{ \begin{array}{c} (1-\alpha)[P(y_x|c) - P(y_{x'}|c)], \\ \alpha[P(y_x|c) - P(y_{x'}|c)], \\ (2\alpha-1)P(y|c) + (1-\alpha)P(y_x|c) - \alpha P(y_{x'}|c), \\ \alpha P(y_x|c) - (1-\alpha)P(y_{x'}|c) - (2\alpha-1)P(y|c) \end{array} \right\} \leq f(\alpha), \tag{A.11}$$

$$min \left\{ \begin{array}{c} \alpha P(y_x|c) - (1-\alpha)P(y_{x'}|c), \\ (1-\alpha)P(y_x|c) + \alpha P(y'_{x'}|c) + \alpha - 1, \\ (2\alpha-1)P(y,x|c) + (2\alpha-1)P(y',x'|c)] + (1-\alpha)[P(y_x|c) - P(y_{x'}|c)], \\ \alpha[P(y_x|c) - P(y_{x'}|c)] + (2\alpha-1)P(y,x'|c) + (2\alpha-1)P(y',x|c) \end{array} \right\} \geq f(\alpha). \tag{A.12}$$

Proof. By lemma 18,

$$\begin{aligned} & f(\alpha) \\ = \; & \alpha P(y_x, y'_{x'}|c) - (1-\alpha)P(y_{x'}, y'_x|c) \\ = \; & \alpha P(y_x, y'_{x'}|c) - (1-\alpha)(P(y_x, y'_{x'}|c) - P(y_x|c) + P(y_{x'}|c)) \\ = \; & (2\alpha-1)P(y_x, y'_{x'}|c) + (1-\alpha)(P(y_x|c) - P(y_{x'}|c)). \end{aligned} \tag{A.13}$$

By lemma 17, substituting (A.1) and (A.2) into (A.13), case 1 and 2 in lemma 19 hold. □

Theorem 4. *The benefit function* $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ *is bounded as follows:*

$$\max\{p_1, p_2, p_3, p_4\} \leq f \leq \min\{p_5, p_6, p_7, p_8\} \text{ if } \sigma < 0,$$
$$\max\{p_5, p_6, p_7, p_8\} \leq f \leq \min\{p_1, p_2, p_3, p_4\} \text{ if } \sigma > 0,$$

where $\sigma, p_1, ..., p_8$ *are given by,*

$$\begin{aligned}
\sigma &= \beta - \gamma - \theta + \delta, \\
p_1 &= (\beta - \theta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
p_2 &= \gamma P(y_x|c) + \delta P(y'_x|c) + (\beta - \gamma)P(y'_{x'}|c), \\
p_3 &= (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta + \delta)[P(y, x|c) + P(y', x'|c)], \\
p_4 &= (\beta - \theta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c) + \\
&\quad (\beta - \gamma - \theta + \delta)[P(y, x'|c) + P(y', x|c)], \\
p_5 &= (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
p_6 &= (\beta - \theta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
p_7 &= (\gamma - \delta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta + \delta)P(y|c), \\
p_8 &= (\beta - \theta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c) - (\beta - \gamma - \theta + \delta)P(y|c).
\end{aligned}$$

Proof.

$$\begin{aligned}
& f(c) \\
= \;& \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y'_x, y_{x'}|c) \\
= \;& \beta P(y_x, y'_{x'}|c) + \gamma[P(y_x|c) - P(y_x, y'_{x'}|c)] + \theta[P(y'_{x'}) - P(y_x, y'_{x'}|c)] + \delta P(y'_x, y_{x'}|c) \\
= \;& \gamma P(y_x|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta)P(y_x, y'_{x'}|c) - (-\delta)P(y'_x, y_{x'}|c). \qquad \text{(A.14)}
\end{aligned}$$

By lemma 19, let $\alpha = \frac{\beta-\gamma-\theta}{\beta-\gamma-\theta-\delta}$, substituting (A.9) to (A.12) into (A.14), theorem 4 hold. □

Theorem 6. *Given that* Y *is monotonic relative to* X*, the benefit function* $f(c)$ *is given by*

$$\begin{aligned}
& f(c) \\
= \;& \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c) \\
= \;& (\beta - \theta)P(y_x|c) + (\gamma - \beta)P(y_{x'}|c) + \theta.
\end{aligned}$$

Proof.

$$
\begin{aligned}
& f(c) \\
= \;& \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y'_x, y_{x'}|c) \\
= \;& \beta[P(y_x|c) - P(y_x, y_{x'}|c)] + \gamma[P(y_{x'}|c) - P(y'_x, y_{x'}|c)] + \\
& \theta[P(y'_x|c) - P(y'_x, y_{x'}|c)] + \delta P(y'_x, y_{x'}|c) \\
= \;& \beta[P(y_x|c) - P(y_{x'}|c) + P(y'_x, y_{x'}|c)] + \gamma[P(y_{x'}|c) - P(y'_x, y_{x'}|c)] + \\
& \theta[P(y'_x|c) - P(y'_x, y_{x'}|c)] + \delta P(y'_x, y_{x'}|c) \\
= \;& \beta P(y_x|c) + (\gamma - \beta)P(y_{x'}|c) + \theta P(y'_x|c) + (\beta + \delta - \gamma - \theta)P(y'_x, y_{x'}|c).
\end{aligned}
$$

Thus, if monotonicity, we have,

$$P(y_{x'}, y'_x|c) = 0. \tag{A.15}$$

Therefore, theorem 6 holds. □

Theorem 8. *Given that the benefit vector $(\beta, \gamma, \theta, \delta)$ satisfies the gain equality, the benefit function $f(c)$ is given by*

$$
\begin{aligned}
& f(c) \\
= \;& \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c) \\
= \;& (\beta - \theta)P(y_x|c) + (\gamma - \beta)P(y_{x'}|c) + \theta.
\end{aligned}
$$

Proof.

$$
\begin{aligned}
& f(c) \\
= \;& \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y'_x, y_{x'}|c) \\
= \;& \beta[P(y_x|c) - P(y_x, y_{x'}|c)] + \gamma[P(y_{x'}|c) - P(y'_x, y_{x'}|c)] + \\
& \theta[P(y'_x|c) - P(y'_x, y_{x'}|c)] + \delta P(y'_x, y_{x'}|c) \\
= \;& \beta[P(y_x|c) - P(y_{x'}|c) + P(y'_x, y_{x'}|c)] + \gamma[P(y_{x'}|c) - P(y'_x, y_{x'}|c)] + \\
& \theta[P(y'_x|c) - P(y'_x, y_{x'}|c)] + \delta P(y'_x, y_{x'}|c) \\
= \;& \beta P(y_x|c) + (\gamma - \beta) P(y_{x'}|c) + \theta P(y'_x|c) + (\beta + \delta - \gamma - \theta) P(y'_x, y_{x'}|c).
\end{aligned}
$$

Thus, with $\beta + \delta = \gamma + \theta$, theorem 8 hold. □

APPENDIX B

Proofs for Chapter 6

Lemma 20. *Given a causal diagram G and distribution compatible with G, let $Z \cup C$ be a set of variables that does not contain any descendant of X in G, then c-specific PNS $P(y_x, y'_{x'}|c)$ is bounded as follows:*

$$\sum_z \max \left\{ \begin{array}{c} 0, \\ P(y_x|z,c) - P(y_{x'}|z,c), \\ P(y|z,c) - P(y_{x'}|z,c), \\ P(y_x|z,c) - P(y|z,c) \end{array} \right\} \times P(z|c) \leq c\text{-}PNS, \tag{B.1}$$

$$\sum_z \min \left\{ \begin{array}{c} P(y_x|z,c), \\ P(y'_{x'}|z,c), \\ P(y,x|z,c) + P(y',x'|z,c), \\ P(y_x|z,c) - P(y_{x'}|z,c) + P(y,x'|z,c) + P(y',x|z,c) \end{array} \right\} \times P(z|c) \geq c\text{-}PNS. \tag{B.2}$$

Proof.

$$\begin{aligned} c\text{-PNS} &= P(y_x, y'_{x'}|c) \\ &= \sum_z P(y_x, y'_{x'}|z,c) \times P(z|c). \end{aligned} \tag{B.3}$$

From Lemma 17, replace c with (z, c), we have the following:

$$\max \left\{ \begin{array}{c} 0, \\ P(y_x|z,c) - P(y_{x'}|z,c), \\ P(y|z,c) - P(y_{x'}|z,c), \\ P(y_x|z,c) - P(y|z,c) \end{array} \right\} \leq P(y_x, y'_{x'}|z,c), \tag{B.4}$$

$$\min \left\{ \begin{array}{c} P(y_x|z,c), \\ P(y'_{x'}|z,c), \\ P(y,x|z,c) + P(y',x'|z,c), \\ P(y_x|z,c) - P(y_{x'}|z,c) + P(y,x'|z,c) + P(y',x|z,c) \end{array} \right\} \geq P(y_x, y'_{x'}|z,c). \tag{B.5}$$

Substituting B.4 and B.5 into B.3, Lemma 20 holds.

Note that since we have,

$$
\begin{aligned}
& \sum_z \max\{0, P(y_x|z,c) - P(y_{x'}|z,c), \\
& P(y|z,c) - P(y_{x'}|z,c), P(y_x|z,c) - P(y|z,c)\} \times P(z|c) \\
\geq & \sum_z 0 \times P(z|c) \\
= & \ 0,
\end{aligned}
$$

$$
\begin{aligned}
& \sum_z \max\{0, P(y_x|z,c) - P(y_{x'}|z,c), \\
& P(y|z,c) - P(y_{x'}|z,c), P(y_x|z,c) - P(y|z,c)\} \times P(z|c) \\
\geq & \sum_z [P(y_x|z,c) - P(y_{x'}|z,c)] \times P(z|c) \\
= & \ P(y_x|c) - P(y_{x'}|c),
\end{aligned}
$$

$$
\begin{aligned}
& \sum_z \max\{0, P(y_x|z,c) - P(y_{x'}|z,c), \\
& P(y|z,c) - P(y_{x'}|z,c), P(y_x|z,c) - P(y|z,c)\} \times P(z|c) \\
\geq & \sum_z [P(y|z,c) - P(y_{x'}|z,c)] \times P(z|c) \\
= & \ P(y|c) - P(y_{x'}|c),
\end{aligned}
$$

$$
\begin{aligned}
& \sum_z \max\{0, P(y_x|z,c) - P(y_{x'}|z,c), \\
& P(y|z,c) - P(y_{x'}|z,c), P(y_x|z,c) - P(y|z,c)\} \times P(z|c) \\
\geq & \sum_z [P(y_x|z,c) - P(y|z,c)] \times P(z|c) \\
= & \ P(y_x|c) - P(y|c),
\end{aligned}
$$

then the lower bound in Lemma 20 is guaranteed to be no worse than the lower bound in Lemma 17. Similarly, the upper bound in Lemma 20 is guaranteed to be no worse than the upper bound in Lemma 17. Also note that, since $Z \cup C$ does not contain a descendant of X, the term $P(y_x|z,c)$ refers to experimental data under population z, c. □

Lemma 21.

$$\begin{aligned} f(c) &= \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c) \\ &= W + \sigma P(y_x, y'_{x'}|c). \end{aligned} \tag{B.6}$$

where,

$$\begin{aligned} W &= (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\ \sigma &= \beta - \gamma - \theta + \delta. \end{aligned}$$

Proof.

$$\begin{aligned} & f(c) \\ = \;& \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y'_x, y_{x'}|c) \\ = \;& \beta P(y_x, y'_{x'}|c) + \gamma [P(y_x|c) - P(y_x, y'_{x'}|c)] + \theta [P(y'_{x'}) - P(y_x, y'_{x'}|c)] + \delta P(y'_x, y_{x'}|c) \\ = \;& \gamma P(y_x|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta) P(y_x, y'_{x'}|c) + \delta P(y'_x, y_{x'}|c). \end{aligned} \tag{B.7}$$

By Lemma 18, we have,

$$P(y'_x, y_{x'}|c) = P(y_x, y'_{x'}|c) - P(y_x|c) + P(y_{x'}|c). \tag{B.8}$$

Substituting B.8 into B.7, we have,

$$\begin{aligned} & f(c) \\ = \;& \gamma P(y_x|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta) P(y_x, y'_{x'}|c) + \delta P(y'_x, y_{x'}|c) \\ = \;& \gamma P(y_x|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta) P(y_x, y'_{x'}|c) + \delta [P(y_x, y'_{x'}|c) - P(y_x|c) + P(y_{x'}|c)] \\ = \;& (\gamma - \delta) P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c) + (\beta - \gamma - \theta + \delta) P(y_x, y'_{x'}|c). \end{aligned}$$

□

Theorem 9. *Given a causal diagram G and distribution compatible with G, let $Z \cup C$ be a set of variables that does not contain any descendant of X in G, then the benefit function $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ is bounded as follows:*

$$W + \sigma U \leq f \leq W + \sigma L \qquad \text{if } \sigma < 0,$$
$$W + \sigma L \leq f \leq W + \sigma U \qquad \text{if } \sigma > 0,$$

where σ, W, L, U are given by,

$$\sigma = \beta - \gamma - \theta + \delta,$$
$$W = (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c),$$
$$L = \sum_z \max \left\{ \begin{array}{c} 0, \\ P(y_x|z,c) - P(y_{x'}|z,c), \\ P(y|z,c) - P(y_{x'}|z,c), \\ P(y_x|z,c) - P(y|z,c) \end{array} \right\} \times P(z|c),$$
$$U = \sum_z \min \left\{ \begin{array}{c} P(y_x|z,c), \\ P(y'_{x'}|z,c), \\ P(y,x|z,c) + P(y',x'|z,c), \\ P(y_x|z,c) - P(y_{x'}|z,c) + P(y,x'|z,c) + P(y',x|z,c) \end{array} \right\} \times P(z|c).$$

Proof. By Lemmas 20 and 21,
substituting B.1 and B.2 into B.6, theorem holds.

Note that, if we substituting Lemma 17 into 21, we have the same results as in Theorem 4. We showed that in Lemma 20 that the bounds in Lemma 20 is guaranteed to be no worse than the bounds in Lemma 17, therefore, the bounds in Theorem 9 is guaranteed to be no worse than the bounds in Theorem 4. □

Lemma 22. *Given a causal diagram G and distribution compatible with G, let $Z \cup C$ be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp\!\!\!\perp X \cup Z_{x'} \mid Z_x, C)$ in G, then the c-PNS*

$P(y_x, y'_{x'}|c)$ is bounded as follows:

$$\max\left\{\begin{array}{c} 0, \\ P(y_x|c) - P(y_{x'}|c), \\ P(y|c) - P(y_{x'}|c), \\ P(y_x|c) - P(y|c) \end{array}\right\} \leq \textit{c-PNS}, \tag{B.9}$$

$$\min\left\{\begin{array}{c} P(y_x|c), \\ P(y'_{x'}|c), \\ P(y, x|c) + P(y', x'|c), \\ P(y_x|c) - P(y_{x'}|c) + P(y, x'|c) + P(y', x|c), \\ \sum_z \sum_{z'} \min\{P(y|z, x, c), P(y'|z', x', c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\} \end{array}\right\} \geq \textit{c-PNS}. \tag{B.10}$$

Proof.

$$\begin{aligned}
& c\text{-PNS} \\
= \; & P(y_x, y'_{x'}|c) \\
= \; & \Sigma_z \Sigma_{z'} P(y_x, y'_{x'}, z_x, z'_{x'}|c) \\
= \; & \Sigma_z \Sigma_{z'} P(y_x, y'_{x'}|z_x, z'_{x'}, c) \times P(z_x, z'_{x'}|c) \\
\leq \; & \Sigma_z \Sigma_{z'} \min\{P(y_x|z_x, z'_{x'}, c), P(y'_{x'}|z_x, z'_{x'}, c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\} \\
= \; & \Sigma_z \Sigma_{z'} \min\{P(y_x|z_x, c), P(y'_{x'}|z'_{x'}, c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\} \qquad \text{(B.11)} \\
= \; & \Sigma_z \Sigma_{z'} \min\{P(y|z_x, x, c), P(y'|z'_{x'}, x', c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\} \qquad \text{(B.12)} \\
= \; & \Sigma_z \Sigma_{z'} \min\{P(y|z, x, c), P(y'|z', x', c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\}.
\end{aligned}$$

Combined with the bounds in Lemma 17, Lemma 22 holds. Note that equation B.11 is due to $Y_x \perp\!\!\!\perp Z_{x'} \mid Z_x, C$ and $Y_{x'} \perp\!\!\!\perp Z_x \mid Z_{x'}, C$. Equation B.12 is due to $\forall x \in X, Y_x \perp\!\!\!\perp X \mid Z_x, C$. □

Theorem 10. *Given a causal diagram G and distribution compatible with G, let Z be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp\!\!\!\perp X \cup Z_{x'} \mid Z_x, C)$ in G, and C does not contain*

any descendant of X in G, then the benefit function $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ is bounded as follows:

$$W + \sigma U \leq f \leq W + \sigma L \qquad \text{if } \sigma < 0,$$
$$W + \sigma L \leq f \leq W + \sigma U \qquad \text{if } \sigma > 0,$$

where σ, W, L, U are given by,

$$\sigma = \beta - \gamma - \theta + \delta,$$
$$W = (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c),$$
$$L = \max \left\{ \begin{array}{c} 0, \\ P(y_x|c) - P(y_{x'}|c), \\ P(y|c) - P(y_{x'}|c), \\ P(y_x|c) - P(y|c) \end{array} \right\},$$
$$U = \min \left\{ \begin{array}{c} P(y_x|c), \\ P(y'_{x'}|c), \\ P(y, x|c) + P(y', x'|c), \\ P(y_x|c) - P(y_{x'}|c) + P(y, x'|c) + P(y', x|c), \\ \sum_z \sum_{z'} \min\{P(y|z, x, c), P(y'|z', x', c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\} \end{array} \right\}.$$

Proof. By Lemmas 22 and 21,
substituting B.9 and B.10 into B.6, theorem holds. □

Lemma 23. *Given a causal diagram G in Figure B.1 and distribution that compatible with G, and C is not a descendant of X, then c-PNS $P(y_x, y'_{x'}|c)$ is bounded as follow:*

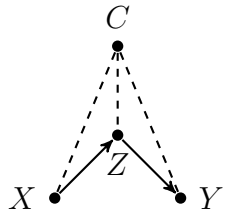

Figure B.1: Mediator Z with no direct effects.

$$\max\left\{\begin{array}{c} 0, \\ P(y_x|c) - P(y_{x'}|c), \\ P(y|c) - P(y_{x'}|c), \\ P(y_x|c) - P(y|c) \end{array}\right\} \leq \textit{c-PNS}, \tag{B.13}$$

$$\min\left\{\begin{array}{c} P(y_x|c), \\ P(y'_{x'}|c), \\ P(y,x|c) + P(y',x'|c), \\ P(y_x|c) - P(y_{x'}|c) + P(y,x'|c) + P(y',x|c), \\ \Sigma_z \Sigma_{z' \neq z} \min\{P(y|z,c), P(y'|z',c)\} \times \min\{P(z|x,c), P(z'|x',c)\} \end{array}\right\} \geq \textit{c-PNS}. \tag{B.14}$$

Proof. First we show that in graph G, if an individual is a c-complier from X to Y, then $Z_x|c$ and $Z_{x'}|c$ must have the different values. This is because the structural equations for Y and Z are $f_y(z, u_y, c)$ and $f_z(x, u_z, c)$, respectively. If an individual has the same $Z_x|c$ and $Z_{x'}|c$ value, then $f_z(x, u_z, c) = f_z(x', u_z, c)$. This means $f_y(f_z(x, u_z, c), u_y, c) = f_y(f_z(x', u_z, c), u_y, c)$, i.e., $Y_x|c$ and $Y_{x'}|c$ must have the same value. Thus this individual is

not a c-complier. Therefore,

$$\begin{aligned}
& c\text{-PNS} \\
= \;& P(y_x, y'_{x'}|c) \\
= \;& \Sigma_z \Sigma_{z' \neq z} P(y_z, y'_{z'}|c) \times P(z_x, z'_{x'}|c) \\
\leq \;& \Sigma_z \Sigma_{z' \neq z} \min\{P(y_z|c), P(y'_{z'}|c)\} \times \min\{P(z_x|c), P(z'_{x'}|c)\} \\
= \;& \Sigma_z \Sigma_{z' \neq z} \min\{P(y|z,c), P(y'|z',c)\} \times \min\{P(z|x,c), P(z'|x',c)\}.
\end{aligned}$$

Combined with the bounds in Lemma 17, Lemma 23 holds. □

Theorem 11. *Given a causal diagram G in Figure B.1 and distribution compatible with G, and C does not contain any descendant of X, then the benefit function $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ is bounded as follows:*

$$\begin{aligned}
W + \sigma U \leq f \leq W + \sigma L \qquad & \text{if } \sigma < 0, \\
W + \sigma L \leq f \leq W + \sigma U \qquad & \text{if } \sigma > 0,
\end{aligned}$$

where σ, W, L, U are given by,

$$\begin{aligned}
& \sigma = \beta - \gamma - \theta + \delta, \\
& W = (\gamma - \delta) P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
& L = \max \left\{ \begin{array}{c} 0, \\ P(y_x|c) - P(y_{x'}|c), \\ P(y|c) - P(y_{x'}|c), \\ P(y_x|c) - P(y|c) \end{array} \right\}, \\
& U = \min \left\{ \begin{array}{c} P(y_x|c), \\ P(y'_{x'}|c), \\ P(y, x|c) + P(y', x'|c), \\ P(y_x|c) - P(y_{x'}|c) + P(y, x'|c) + P(y', x|c), \\ \Sigma_z \Sigma_{z' \neq z} \min\{P(y|z,c), P(y'|z',c)\} \times \min\{P(z|x,c), P(z'|x',c)\} \end{array} \right\}.
\end{aligned}$$

Proof. By Lemmas 23 and 21,

substituting B.13 and B.14 into B.6, theorem holds. □

APPENDIX C

Proofs for Chapter 7

Theorem 12. *The benefit function* $f(c) = \beta P(y_x, y'_{x'}|c) + \gamma P(y_x, y_{x'}|c) + \theta P(y'_x, y'_{x'}|c) + \delta P(y_{x'}, y'_x|c)$ *is bounded as follows:*

$$max\{p_1, p_2\} \leq f \leq min\{p_3, p_4\} \text{ if } \sigma < 0, \tag{C.1}$$

$$max\{p_3, p_4\} \leq f \leq min\{p_1, p_2\} \text{ if } \sigma > 0, \tag{C.2}$$

where $\sigma, p_1, ..., p_4$ *are given by,*

$$\begin{aligned}
\sigma &= \beta - \gamma - \theta + \delta, \\
p_1 &= (\beta - \theta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
p_2 &= \gamma P(y_x|c) + \delta P(y'_x|c) + (\beta - \gamma)P(y'_{x'}|c), \\
p_3 &= (\gamma - \delta)P(y_x|c) + \delta P(y_{x'}|c) + \theta P(y'_{x'}|c), \\
p_4 &= (\beta - \theta)P(y_x|c) - (\beta - \gamma - \theta)P(y_{x'}|c) + \theta P(y'_{x'}|c).
\end{aligned}$$

Proof. Theorem 12 directly follows Theorem 4 by removing all $P(X, Y|c)$ and $P(Y|c)$ parts. Because the L.H.S. of the Equations C.1 and C.2 are max functions, and the R.H.S. are min functions, therefore, we loose the bounds by removing terms, and thus the bounds are still vaild. □

Theorem 13. *Given a causal diagram* G *and a distribution compatible with* G*, let* $W \cup U$ *be a set of variables satisfying the back-door criterion in* G *relative to an ordered pair* (X, Y)*, where* $W \cup U$ *is partially observable, i.e., only probabilities* $P(X, Y, W)$ *and* $P(U)$ *are given,*

the causal effects of X on Y are then bounded as follows:

$$LB \leq P(y|do(x)) \leq UB$$

where LB is the solution to the non-linear optimization problem in Equation C.3 and UB is the solution to the non-linear optimization problem in Equation C.4.

$$LB = \min \sum_{w,u} \frac{a_{w,u} b_{w,u}}{c_{w,u}}, \tag{C.3}$$

$$UB = \max \sum_{w,u} \frac{a_{w,u} b_{w,u}}{c_{w,u}}, \tag{C.4}$$

where,

$$\sum_u a_{w,u} = P(x, y, w), \sum_u b_{w,u} = P(w), \sum_u c_{w,u} = P(x, w) \text{ for all } w \in W;$$

and for all $w \in W$ and $u \in U$,

$$b_{w,u} \geq c_{w,u} \geq a_{w,u},$$

$$\max\{0, p(x, y, w) + p(u) - 1\} \leq a_{w,u} \leq \min\{P(x, y, w), p(u)\},$$

$$\max\{0, p(w) + p(u) - 1\} \leq b_{w,u} \leq \min\{P(w), p(u)\},$$

$$\max\{0, p(x, w) + p(u) - 1\} \leq c_{w,u} \leq \min\{P(x, w), p(u)\}.$$

Proof. To show that the non-linear optimization bounds the actual causal effects, we only need to show that there exists a point in feasible space that $\sum_{w,u} \frac{a_{w,u} b_{w,u}}{c_{w,u}}$ is equal to the actual causal effects.

Since $W \cup U$ satisfies the back-door criterion, by adjustment formula in Equation 3.1, we have,

$$\begin{aligned} P(y|do(x)) &= \sum_{w,u} P(y|x, w, u) P(w, u) \\ &= \sum_{w,u} \frac{P(x, y, w, u) P(w, u)}{P(x, w, u)}. \end{aligned}$$

Let

$$a_{w,u} = P(x, y, w, u),$$
$$b_{w,u} = P(w, u),$$
$$c_{w,u} = P(x, w, u).$$

We now show that the above set of $a_{w,u}, b_{w,u}, c_{w,u}$ are in feasible space.
We have,

$$\text{for } w \in W,$$
$$\sum_u a_{w,u} = \sum_u P(x, y, w, u) = P(x, y, w),$$
$$\sum_u b_{w,u} = \sum_u P(w, u) = P(w),$$
$$\sum_u c_{w,u} = \sum_u P(x, w, u) = P(x, w),$$

and for all $w \in W$ and $u \in U$,

$$b_{w,u} = P(w, u) \geq P(x, w, u) = c_{w,u},$$
$$c_{w,u} = P(x, w, u) \geq P(x, y, w, u) = a_{w,u},$$
$$a_{w,u} = P(x, y, w, u) \leq \min\{P(x, y, w), p(u)\},$$
$$b_{w,u} = P(w, u) \leq \min\{P(w), p(u)\},$$
$$c_{w,u} = P(x, w, u) \leq \min\{P(x, w), p(u)\},$$
$$a_{w,u} = P(x, y, w, u) \geq \max\{0, p(x, y, w) + p(u) - 1\},$$
$$b_{w,u} = P(w, u) \geq \max\{0, p(w) + p(u) - 1\},$$
$$c_{w,u} = P(x, w, u) \geq \max\{0, p(x, w) + p(u) - 1\}.$$

Therefore, the above set of $a_{w,u}, b_{w,u}, c_{w,u}$ are in feasible space, and thus, the UB and LB bound the actual causal effects. □

Theorem 14. *Given a causal diagram G and distribution compatible with G, let $W \cup U$ be a set of variables satisfying the front-door criterion in G relative to an ordered pair (X, Y),*

where $W \cup U$ is partially observable, i.e., only probabilities $P(X, Y, W)$ and $P(U)$ are given and $P(x, W, U) > 0$, the causal effects of X on Y are then bounded as follows:

$$LB \leq P(y|do(x)) \leq UB$$

where LB is the solution to the non-linear optimization problem in Equation C.5 and UB is the solution to the non-linear optimization problem in Equation C.6.

$$LB = \min \sum_{w,u} \frac{b_{x,w,u}}{P(x)} \sum_{x'} \frac{a_{x',w,u} P(x')}{b_{x',w,u}}, \tag{C.5}$$

$$UB = \max \sum_{w,u} \frac{b_{x,w,u}}{P(x)} \sum_{x'} \frac{a_{x',w,u} P(x')}{b_{x',w,u}}, \tag{C.6}$$

where,

$\sum_u a_{x,w,u} = P(x, y, w), \sum_u b_{x,w,u} = P(x, w)$ *for all* $x \in X$ *and* $w \in W$;
and for all $x \in X$,$w \in W$, *and* $u \in U$,

$b_{x,w,u} \geq a_{x,w,u}$,

$\max\{0, p(x, y, w) + p(u) - 1\} \leq a_{x,w,u} \leq \min\{P(x, y, w), p(u)\}$,

$\max\{0, p(x, w) + p(u) - 1\} \leq b_{x,w,u} \leq \min\{P(x, w), p(u)\}$.

Proof. To show that the non-linear optimization bounds the actual causal effects, we only need to show that there exists a point in feasible space that $\sum_{w,u} \frac{b_{x,w,u}}{P(x)} \sum_{x'} \frac{a_{x',w,u} P(x')}{b_{x',w,u}}$ is equal to the actual causal effects.

Since $W \cup U$ satisfies front-door criterion and $P(u, W, U) > 0$, by adjustment formula in Equation 3.2, we have,

$$\begin{aligned} P(y|do(x)) &= \sum_{w,u} P(w, u|x) \sum_{x'} P(y|x', w, u) P(x') \\ &= \sum_{w,u} \frac{P(x, w, u)}{P(x)} \sum_{x'} \frac{P(x', y, w, u) P(x')}{P(x', w, u)}. \end{aligned}$$

Let

$$a_{x,w,u} = P(x, y, w, u),$$

$$b_{x,w,u} = P(x, w, u).$$

Similarly to the proof of Theorem 13, it is easy to show that the above set of $a_{x,w,u}, b_{x,w,u}$ are in feasible space, and therefore, LB and UB bound the actual causal effects. □

Theorem 16. *Let G be a causal diagram containing nodes $\{V_1, ..., V_{n-3}, X, Y, Z\}$. Let O be any observational data compatible with G. Suppose there exists a set of variables that satisfies the back-door or front-door criterion relative to (X, Y) in G, then, (G, O) is equivalent to (G', O') (G' containing nodes $\{V_1, ..., V_{n-3}, X, Y, W, U\}$; O' is observational data compatible with G'), where the number of states in W times the number of states in U is equal to the number of states in Z, and the structure of G' and the observational data O' are obtained as follows:*

Structure of G':

Let $Parents_G(H)$ be the parents of H in causal diagram G.

$Parents_{G'}(U) = Parents_G(Z)$,

$Parents_{G'}(W) = Parents_G(Z) \cup \{U\}$.

For $H \in \{V_1, ..., V_{n-3}, X, Y\}$,

$Parents_{G'}(H) = Parents_G(H)$ *if* $Z \notin Parents_G(H)$,

$Parents_{G'}(H) = Parents_G(H) \setminus \{Z\} \cup \{W, U\}$ *if* $Z \in Parents_G(H)$.

Note that, let Q be the set of variables in G that satisfies the back-door or front-door criterion relative to (X, Y), then Q' satisfies the back-door or front-door criterion relative to (X, Y) in G' , where

$Q' = Q$ *if* $Z \notin Q$,

$Q' = Q \setminus \{Z\} \cup \{W, U\}$ *if* $Z \in Q$.

Observational data:

Let p be the number of states in W, and let q be the number of states in U.

The states of Z are the Cartesian product of the states of W and the states of U.

In detail,

(w_j, u_k) *is equivalent to* $z_{(j-1)*q+k}$,

w_j *is equivalent to* $\vee_{k=1}^{q}(w_j, u_k) = \vee_{k=1}^{q} z_{(j-1)*q+k}$,

u_k is equivalent to $\vee_{j=1}^{p}(w_j, u_k) = \vee_{j=1}^{p} z_{(j-1)*q+k}$,
$P(w_j, u_k, V) = P(z_{(j-1)*q+k}, V)$ for any $V \subseteq \{V_1, ..., V_{n-3}, X, Y\}$.

Proof. First, we show that Q' satisfies the back-door or front-door criterion relative to (X, Y) in G'. If Q satisfies the back-door criterion relative to (X, Y) in G, we need to show that,

- no node in Q' is a descendant of X.
- Q' blocks every path between X and Y that contains an arrow into X.

It is easy to show that if there is a node in Q' that is a descendant of X in G', then there is a node in Q that is a descendant of X in G. And if there is a path between X and Y that contains an arrow into X does not blocked by Q' in G', then there is a path between X and Y that contains an arrow into X does not blocked by Q in G. Thus, Q' satisfies the back-door criterion relative to (X, Y) in G'. Similarly, we can show that if Q satisfies the front-door criterion relative to (X, Y) in G, then Q' satisfies the front-door criterion relative to (X, Y) in G'.

Now, we show that (G, O) is equivalent to (G', O'), i.e., show that $P(y|do(x))$ is the same between (G, O) and (G', O'). Suppose Q satisfies the back-door criterion relative to (X, Y) in G. By adjustment formula in Equation 3.1, we have,
$P(y|do(x)) = \sum_{q \in Q} P(y|do(x), q) \times P(q)$.
And in G',
$P(y|do(x)) = \sum_{q \in Q'} P(y|do(x), q) \times P(q)$,
it is obviously that these two causal effects are the same,
because $P(w_j, u_k, V) = P(z_{(j-1)*q+k}, V)$ for any $V \subseteq \{V_1, ..., V_{n-3}, X, Y\}$.
Similarly, we can show that if Q satisfies the front-door criterion relative to (X, Y) in G, (G, O) is equivalent to (G', O'). □

REFERENCES

[AIR96] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association*, **91**(434):444–455, 1996.

[BCS01] Marsha Blumenthal, Charles Christian, Joel Slemrod, and Matthew G Smith. "Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota." *National Tax Journal*, pp. 125–138, 2001.

[BP97a] Alexander Balke and Judea Pearl. "Bounds on treatment effects from studies with imperfect compliance." *Journal of the American Statistical Association*, **92**(439):1171–1176, 1997.

[BP97b] Alexander A Balke and Judea Pearl. "Probabilistic counterfactuals: Semantics, computation, and applications." Technical report, Department of Computer Science, UCLA, 1997.

[BPQ13] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. "Counterfactual reasoning and learning systems: The example of computational advertising." *The Journal of Machine Learning Research*, **14**(1):3207–3260, 2013.

[BST99] Alex Berson, Stephen Smith, and Kurt Thearling. *Building data mining applications for CRM*. McGraw-Hill Professional, 1999.

[GP98] David Galles and Judea Pearl. "An axiomatic characterization of causal counterfactuals." *Foundations of Science*, **3**(1):151–182, 1998.

[Hal00] Joseph Y Halpern. "Axiomatizing causal reasoning." *Journal of Artificial Intelligence Research*, **12**:317–337, 2000.

[HYW06] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management." *Expert Systems with Applications*, **31**(3):515–524, 2006.

[KC11] Manabu Kuroki and Zhihong Cai. "Statistical Analysis of 'Probabilities of Causation' Using Co-variate Information." *Scandinavian Journal of Statistics*, **38**(3):564–577, 2011.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: Principles and techniques*. MIT press, 2009.

[Kra88] Dieter Kraft. *A software package for sequential quadratic programming*. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht. Wiss. Berichtswesen d. DFVLR, 1988.

[LCK14] Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. "Counterfactual estimation and optimization of click metrics for search engines." *arXiv preprint arXiv:1403.1891*, 2014.

[Lej01] Miguel APM Lejeune. "Measuring the impact of data mining on churn management." *Internet Research*, **11**(5):375–387, 2001.

[LR14] Randall A Lewis and David H Reiley. "Online ads and offline sales: Measuring the effect of retail advertising via a controlled experiment on Yahoo!" *Quantitative Marketing and Economics*, **12**(3):235–266, 2014.

[MLP21] Scott Mueller, Ang Li, and Judea Pearl. "Causes of Effects: Learning individual responses from population data." *arXiv preprint arXiv:2104.13730*, 2021.

[Pea93] J Pearl. "Aspects of Graphical Models Connected With Causality." *Proceedings of the 49th Session of the international Statistical Institute, Italy*, pp. 399–401, 1993.

[Pea95] Judea Pearl. "Causal diagrams for empirical research." *Biometrika*, **82**(4):669–688, 1995.

[Pea99] Judea Pearl. "Probabilities of causation: Three counterfactual interpretations and their identification." *Synthese*, **121**(1-2):93–149, 1999.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Pea14] Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 2014.

[Ros75] John T. Roscoe. *Fundamental Research Statistics for the Behavioral Sciences*. Number v. 2 in Editors' Series in Marketing. Holt, Rinehart and Winston, 1975.

[RZS06] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. "The value of reputation on eBay: A controlled experiment." *Experimental economics*, **9**(2):79–101, 2006.

[Sci20] SciPyCommunity. "Scipy Reference Guide.", 2020.

[SGS00] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[SNO98] S Shyam Sundar, Sunetra Narayan, Rafael Obregon, and Charu Uppal. "Does web advertising work? Memory for print vs. online media." *Journalism & Mass Communication Quarterly*, **75**(4):822–835, 1998.

[SWY15] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. "Causal Inference via Sparse Additive Models with Application to Online Advertising." In *AAAI*, pp. 297–303, 2015.

[TL09] Chih-Fong Tsai and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." *Expert Systems with Applications*, **36**(10):12547–12553, 2009.

[TP00] Jin Tian and Judea Pearl. "Probabilities of causation: Bounds and identification." *Annals of Mathematics and Artificial Intelligence*, **28**(1-4):287–313, 2000.

[Win01] Russell S Winer. "A framework for customer relationship management." *California management review*, **43**(4):89–105, 2001.

[YLW09] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. "How much can behavioral targeting help online advertising?" In *Proceedings of the 18th international conference on World wide web*, pp. 261–270. ACM, 2009.