Department of Computer Engineering
TUM School of Computation, Information and Technology
Technical University of Munich

Multi-objective causal Bayesian optimisation

Shriya Bhatija

Thesis for the attainment of the academic degree

M.Sc. Mathematics

at the TUM School of Computation, Information and Technology of the Technical University of Munich

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisors:
M.Sc. Jakob Thumm, M.Sc. Paul-David Zuercher (external), Dr. Thomas Bohné (external)

Submitted:
Munich, December 15, 2024

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, December 15, 2024 Shriya Bhatija

Abstract

We propose *multi-objective causal Bayesian optimisation* (MO-CBO), a new problem class for identifying Pareto-optimal interventions that simultaneously optimise multiple target variables within a known causal graph. MO-CBO extends the *causal Bayesian optimisation* (CBO) family of methods to support optimisation on causal models with multiple outcomes. We prove that any MO-CBO problem can be decomposed into a series of traditional multi-objective optimisation tasks, and introduce CAUSAL PARETOSELECT, an algorithm that sequentially balances exploration across these tasks using relative hypervolume improvement. Our methodology generalises multi-objective Bayesian optimisation to perform causally-informed function evaluations, instead of neglecting known causal relationships. By establishing graphical criteria, we enforce CAUSAL PARETOSELECT to explore only potentially optimal sets of variables to intervene upon. We validate our approach on both synthetic and real-world causal graphs, demonstrating its superiority over traditional multi-objective Bayesian optimisation.

Contents

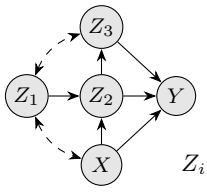
1 Introduction

Decision making problems arise in a variety of domains, such as healthcare, manufacturing or public policy, and involve implementing interventions and manipulating variables in order to obtain an outcome of interest (Pearl & Mackenzie, 2018). For instance, in drug design, the goal is to derive drug formulations that allow for targeted therapy while minimising side effects, ensuring both efficacy and patient safety. Such systems are often governed by causal mechanisms, and accounting for their underlying relationships is crucial for making informed decisions (Imbens & Rubin, 2015). In real-world settings, simply performing large-scale interventions without understanding the data generating process, can be an economically inviable strategy for attaining a desired outcome (Imbens & Rubin, 2015). Interventions are inherently costly, and practical applications are often subject to budgetary constraints. Hence, it is favourable to exploit the known causal relationships, and identify optimal solutions in a targeted and cost-efficient manner.

In these data-scarce and cost-sensitive settings, Bayesian optimisation has shown to be an efficient approach (Shahriari et al., 2016). Aglietti et al., 2020 introduce *causal Bayesian optimisation* (CBO), a method that generalises Bayesian optimisation to enforce causally-informed function evaluations. While existing CBO variants focus on optimising a single objective, real-world systems are complex and may require simultaneous optimisation of multiple outcomes (Stewart et al., 2008). To this end, we propose a new paradigm called *multi-objective causal Bayesian optimisation* (MO-CBO) and derive an algorithm, called CAUSAL PARETOSELECT, specifically designed for multi-objective optimisation in causal systems. We demonstrate that our method outperforms traditional multi-objective Bayesian optimisation in settings where causal information is available, achieving more cost-effective, diverse, and accurate solutions.

1.1 Problem statement and contributions

We consider the problem of optimising an output variable within a causal system in which a sequence of interventions can be performed. Here, so called *structural causal models* are widely used for probabilistic causal modelling as they come with a coherent mathematical formalism for understanding and predicting causal effects under uncertainty (Pearl, 2000). In our problem setting, we assume the graph topology to be known, while the precise functional mechanisms between the variables remain unknown. Focusing on a specific example, consider



Y: crop yield
X: soil fumigants
Z_i: eel-worm population at time i, $i = 1, 2, 3$

Figure 1.1: Crop yield example (Cochran & Cox, 1957). Nodes are variables, arrows show causal effects, and dashed edges indicate unobserved common causes.

the causal graph given in Figure 1.1. Here, Y represents crop yields for various agricultural products, X denotes soil fumigants, and $\mathbf{Z} = \{Z_1, Z_2, Z_3\}$ represents the eel-worm population at different times (Cochran & Cox, 1957). The goal is to identify the most effective value to which X shall be set to (intervention) in order to maximise the crop yield Y, within a limited number of seasons and subject to a budget constraint, as each intervention incurs a specific cost. For optimisation problems of this type, Aglietti et al., 2020 introduce *causal Bayesian optimisation* (CBO) enabling the incorporation of causal information in the Bayesian optimisation process. More specifically, the CBO algorithm optimises in a two fold manner: selecting a set of variables to be intervened upon, called *intervention set*, along with the values to which these variables shall be set to, referred to as *intervention values*. Thus, CBO aims to find an *intervention set-value* pair that optimises an outcome variable of interest. Throughout several experiments, the authors demonstrate, that their approach is cost-effective and reduces the number of function evaluations as compared to traditional Bayesian optimisation.

To the best of our knowledge, a multi-objective variant of CBO has not yet been proposed, and *multi-objective causal Bayesian optimisation* (MO-CBO) shall be the contribution of this thesis. For instance, Figure 1.1 could be extended to a multi-objective system if the various crop yields were modelled as distinct outcomes $Y_1, \ldots, Y_m$. The associated MO-CBO problem is then the simultaneous maximisation of all Y_i, while leveraging the known causal relationships between the variables. Generally, in multi-objective optimisation, conflicting objectives often result in a set of *Pareto-optimal* solutions rather than a single global optimum (Miettinen, 1999). In MO-CBO, the aim is to find Pareto-optimal solutions in terms of intervention set-value pairs.

The key contributions of this thesis are as follows:

1. We provide a literature review for embedding our research within the broader context of related fields, such as multi-objective Bayesian optimisation, causality theory, and causal decision-making.

2. We establish a theoretical foundation for our approach by extending an existing methodology, derived by Lee and Bareinboim, 2018, to the multi-objective setting. The method offers a mathematical justification for choosing intervention sets in order to optimise given outcomes variables. It accounts for the presence of unobserved common causes, which can misleadingly suggest causal relationships between variables.

3. We formally propose *multi-objective causal Bayesian optimisation* (MO-CBO) as a new class of optimisation problems, for which we define the notion of *causal Pareto-optimality.*

4. We develop an algorithm, called Causal ParetoSelect, that balances the exploration of multiple intervention sets. This method builds upon our extension of the work from Lee and Bareinboim, 2018.

5. Through an experimental study, we demonstrate that our approach surpasses traditional multi-objective Bayesian optimisation in settings where causal relationships are known. Specifically,

- Causal ParetoSelect reaches Pareto-optimal solutions with fewer interventions, offering a more cost-efficient approach.
- Causal ParetoSelect uncovers a more diverse range of Pareto-optimal solutions.
- Causal ParetoSelect identifies solutions that are not attainable by traditional multi-objective Bayesian optimisation.

1.2 Thesis structure

Continuing with chapter 2, we provide a concise overview of the theoretical concepts relevant for our work. In chapter 3, we present a literature review to contextualise our research within already established fields. Thereafter, in chapter 4, we offer theoretical contributions on selecting feasible intervention sets in multi-objective structural causal models. These contributions lay the groundwork for chapter 5, where we formally define the problem class of mo-cbo and develop the Causal ParetoSelect algorithm for solving mo-cbo problems. Subsequently, the applicability of our method will be demonstrated in chapter 6 through a series of experiments in both synthetic and real-world settings. In chapter 7, we reflect on our methodology, addressing its limitations and exploring potential directions for future work. Finally, the thesis concludes with a review of the main insights and contributions in chapter 8.

2 Theoretical concepts

This chapter provides a concise overview of the theoretical background relevant for multi-objective causal Bayesian optimisation (MO-CBO), the novel problem formulation proposed in this thesis. We begin in section 2.1 by reviewing the core principles of Bayesian optimisation and its extension to the multi-objective setting. In section 2.2, the focus shifts to concepts in causality theory. We introduce so called *structural causal models*, which are a standard tool for representing and analysing causal systems. Thereafter, we formally define the concept of manipulating variables within these causal models, encompassing a mathematical description of *interventions* and their causal effect propagation. Finally, we briefly explore *do-calculus*, which is an axiomatic system for calculating interventional effects and is regarded invaluable in the field of causality. The concepts introduced in this chapter provide a theoretical foundation and justification for the methodologies adopted in our work. The proposed notation will be used throughout the thesis.

2.1 Bayesian optimisation

In this section, we review the Bayesian optimisation procedure. Unless otherwise stated, the explanations are either based on the review of Shahriari et al., 2016 or, for the multi-objective variant, on the works of Miettinen, 1999 and Konakovic Lukovic et al., 2020.

2.1.1 Methodology

Bayesian optimisation is a methodology for globally optimising functions that are accessible only through input-output observations, commonly referred to as *black-box functions*. Formally, the aim is to minimise (or, for the dual formulation, maximise)

$$x^* = \arg\min_{x \in \mathcal{X}} f(x), \tag{2.1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is the objective function with input/design space $\mathcal{X}$ (Shahriari et al., 2016). Bayesian optimisation is designed to be sample-efficient, making it particularly valuable when function evaluations are expensive to perform, and little is known about the objective's analytical from. The core principle is to iteratively learn from previous evaluations to infer regions in the design space probable to contain the optimum. Generally, each iteration of a Bayesian optimisation algorithm proceeds as follows: The existing data points, comprising prior inputs $x_1, \ldots, x_{n-1}$ and their corresponding evaluations $y_1, \ldots, y_{n-1}$, are used to construct an approximate statistical model $\tilde{f} : \mathcal{X} \to \mathbb{R}$ of the objective function, known as the *surrogate model*. Based on $\tilde{f}$, an inexpensive *acquisition function* $\alpha : \mathcal{X} \to \mathbb{R}_+$ estimates the utility of evaluating

Algorithm 1: Bayesian optimisation

Input: Design space $\mathcal{X}$, objective function f, initial samples $\mathcal{D}_0$
Output: Global optimum x^* and $f(x^*)$
for $n = 1, \ldots, N$ **do**
  Train a surrogate model $\tilde{f}$ using $\mathcal{D}_{n-1}$;
  Select the next point, i.e. $x_n = \arg\max_{x \in \mathcal{X}} \ \alpha(\tilde{f}(x), \mathcal{D}_{n-1})$;
  Evaluate the objective function, i.e. $y_n = f(x_n)$;
  Augment the dataset, i.e. $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{(x_n, y_n)\}$;
Compute the global optimum x^* and $f(x^*)$ from $\mathcal{D}_N$

the objective function at a given point. The next point to be evaluated, x_n, is then determined by maximising the acquisition function:

$$x_n = \arg\max_{x \in \mathcal{X}} \ \alpha(\tilde{f}(x), \mathcal{D}_{n-1}). \tag{2.2}$$

This process is repeated until a specified budget is exhausted or the optimum is identified. For a concise presentation of the procedure, we refer to algorithm 1. While Bayesian optimisation is mathematically more complex than for instance grid or random search, its reduced reliance on objective function evaluations makes it an efficient method for data-scarce and expensive-to-evaluate scenarios. It is designed to balance between *exploration* (gathering information in uncertain regions) and *exploitation* (focusing on areas likely to yield the optimum based on previous data) (Shahriari et al., 2016). In the following paragraphs, we will discuss the most common surrogate model and acquisition function.

Gaussian processes. Gaussian processes (Rasmussen, 2003) are widely employed as surrogate models in Bayesian optimisation due to their ability to probabilistically model unknown objectives by defining distributions them. We state the the formal definition as well as briefly discuss Gaussian process regression.

Definition 2.1 (Gaussian process)**.** A *Gaussian process* is a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions.

A Gaussian process is fully specified by its mean function $m(x)$ and covariance function $k(x, x')$, forming a natural generalisation of the Gaussian distribution, where the mean and covariance are a vector and a matrix, respectively. The Gaussian distribution is defined over vectors, whereas a Gaussian process is defined over functions. We denote

$$f \sim \mathcal{GP}(m, k) \tag{2.3}$$

for a function $f : \mathcal{X} \to \mathbb{R}$ that is distributed as a Gaussian Process with mean function m and covariance function k. Focusing on a specific function evaluation, we can also write $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$ to denote the mean and uncertainty of f evaluated at x.

Gaussian process regression (Rasmussen, 2003) is a technique for modelling black-box objectives by sampling data points and fitting a mean and covariance function. It is a non-parametric regression approach where an unknown function can be estimated as an instance of a Gaussian process. The so called *prior function* reflects the assumptions about the objective and can already incorporate prior data if available. After observing a set of input-output pairs, Gaussian process regression updates its beliefs about the objective, producing a *posterior distribution* that provides both predictions (mean function) and uncertainty estimates (covariance). For the precise derivation of Gaussian process regression, we refer to Rasmussen, 2003.

Expected improvement. The acquisition function balances the exploration-exploitation trade-off in Bayesian optimisation as it drives the selection of the next evaluation point (Shahriari et al., 2016). One of the most popular acquisition functions is the *expected improvement* (Jones et al., 1998), that operates as follows: At each iteration, it estimates the expected amount of improvement over the current best observation $f(x_{opt})$ in $\mathcal{D}_{n-1}$, given the surrogate evaluation $\tilde{f}(x)$ of a new input x. The value of the optimum is then either $f(x_{opt})$ or $\tilde{f}(x)$, and the improvement is given by $(f(x_{opt}) - \tilde{f}(x))^+$, where $a^+ = \max(a, 0)$. Thus, the next point to be evaluated, x_n, is chosen to maximise the expected improvement,

$$x_n = \arg\max_{x \in \mathcal{X}} \left(\mathbb{E}[f(x_{opt}) - \tilde{f}(x)] \right)^+, \tag{2.4}$$

where $\tilde{f}(x)$ is a Gaussian random variable as previously discussed.

2.1.2 Multi-objective extension (MOBO)

Multi-objective Bayesian optimisation (MOBO) extends the principles of Bayesian optimisation to scenarios involving several objective functions. Rather than optimising a single black-box function, the aim is to simultaneously minimise (or, in the dual case, maximise) a collection of functions $f_1, \ldots, f_m : \mathcal{X} \to \mathbb{R}$, where each function represents a different objective. The challenge in optimising multiple objectives arises from their possibly conflicting behaviour, as the optimal regions may not coincide across all functions. Consequently, the solution to such a problem is not a single global optimum, but rather a set of solutions known as *Pareto front* (Miettinen, 1999). The following definitions, taken from (Miettinen, 1999), formalise this concept.

Definition 2.2 (Dominance)**.** Given objectives $f_1, \ldots, f_m : \mathcal{X} \to \mathbb{R}$, a point $x \in \mathcal{X}$ *dominates* another point $x' \in \mathcal{X}$ if it holds $f_i(x') \geqslant f_i(x)$ for all $1 \leqslant i \leqslant m$, and $f_i(x') > f_i(x)$ for at least one $1 \leqslant i \leqslant m$. This relation is denoted as $x \prec x'$. If equality is permitted in all components, we say x *weakly dominates* x', denoted $x \preceq x'$.

Definition 2.3 (Pareto-optimality)**.** Given objectives $f_1, \ldots, f_m : \mathcal{X} \to \mathbb{R}$, a point $x \in \mathcal{X}$ is called *Pareto-optimal* or *non-dominated* if there is no other $x' \in \mathcal{X}$ such that $f_i(x) \geqslant f_i(x')$ for all $1 \leqslant i \leqslant m$, and $f_i(x) > f_i(x')$ for at least one $1 \leqslant i \leqslant m$. The set of Pareto-optimal points in $\mathcal{X}$ is called *Pareto set* and will be denoted $\mathcal{P}_s$. The *Pareto front* is the image of the Pareto set under the objective functions, given by $\mathcal{P}_f = \{(f_1(x), \ldots, f_m(x)) \mid x \in \mathcal{P}_s\}$.

Algorithm 2: Multi-objective Bayesian optimisation (MOBO)

Input: Design space $\mathcal{X}$, objective functions $f_1, \ldots, f_m$, initial samples $\mathcal{D}_0$
Output: Pareto set $\mathcal{P}_s$ and Pareto front $\mathcal{P}_f$
for $n = 1, \ldots, N$ **do**
  Fit surrogate models for $\tilde{f}_1, \ldots, \tilde{f}_m$ using $\mathcal{D}_{n-1}$;
  Approximate Pareto set $\tilde{\mathcal{P}}_s$ and Pareto front $\tilde{\mathcal{P}}_f$ based on the surrogates $\tilde{f}_i$;
  Select next point, i.e. $x_n = \arg\max_{x \in \mathcal{X}} \alpha(x, \tilde{\mathcal{P}}_s, \tilde{\mathcal{P}}_f)$;
  Evaluate objective functions, i.e. $\mathbf{y}_n = (f_1(x_n), \ldots, f_m(x_n))$;
  Augment the dataset, i.e. $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{(x_n, \mathbf{y}_n)\}$;
Compute Pareto front $\mathcal{P}_f$ and Pareto set $\mathcal{P}_s$ from $\mathcal{D}_N$

In other words, a point $x \in \mathcal{X}$ is Pareto-optimal if improving any one objective function necessarily results in worsening at least one other objective. The *Pareto front* establishes *non-dominated trade-offs* between the functions, and is therefore the solution to a multi-objective optimisation problem. Thus, a MOBO algorithm aims to approximate the Pareto front while minimising the number of function evaluations. Typically, independent Gaussian processes are employed to model each objective. In every iteration, after updating the surrogate model, an approximation of the Pareto set and Pareto front is computed (Konakovic Lukovic et al., 2020). The complete procedure is given in algorithm 2.

Hypervolume improvement. In the multi-objective setting, the acquisition function needs to measure the quality of an approximated Pareto front. To this end, the so called *hypervolume* indicator (Zitzler & Thiele, 1999) is the most commonly used metric. Let $\tilde{\mathcal{P}}_f$ be a Pareto front approximation and $\mathbf{r} \in \mathbb{R}^m$ a reference point. Typically, $\mathbf{r}$ is chosen to be worse than or equal to the worst possible value for each objective. The hypervolume is defined as

$$\mathcal{H}(\tilde{P}_f) = \int_{\mathbb{R}^m} \mathbb{1}_{H(\tilde{P}_f)}(z) dz, \tag{2.5}$$

where $H(\tilde{P}_f) = \{\mathbf{z} \in \mathbb{R}^m \mid \exists 1 \leqslant i \leqslant |\tilde{\mathcal{P}}_f| : \mathbf{r} \preceq \mathbf{z} \preceq \tilde{\mathcal{P}}_f(i)\}$ and $\tilde{\mathcal{P}}_f(i)$ denotes the i-th solution in $\tilde{\mathcal{P}}_f$. Additionally, $\mathbb{1}_{H(\tilde{P}_f)}$ represents the Dirac delta function that equals 1 if $\mathbf{z} \in H(\tilde{\mathcal{P}}_f)$, and 0 otherwise. The higher the hypervolume, the better $\tilde{\mathcal{P}}_f$ approximates the true Pareto front. To evaluate the potential improvement in hypervolume when adding a set of new points $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_k\} \subset \mathbb{R}^m$ to the current Pareto front approximation, the *hypervolume improvement* is defined as

$$\mathrm{HVI}(\mathbf{P}, \tilde{\mathcal{P}}_f) = \mathcal{H}(\tilde{\mathcal{P}}_f \cup \mathbf{P}) - \mathcal{H}(\tilde{\mathcal{P}}_f). \tag{2.6}$$

Therefore, in each iteration, MOBO selects the next point to be evaluated by maximising the hypervolume improvement. In batched MOBO approaches, a set of points is chosen for evaluation in parallel, where the hypervolume improvement selects a batch of points as

$$x_{bn+1}, \ldots, x_{(b+1)n} = \underset{\mathbf{P} \subseteq \mathcal{X},\ |\mathbf{P}| = b}{\arg\max} \mathrm{HVI}(\mathbf{P}, \tilde{\mathcal{P}}_f), \tag{2.7}$$

where b denotes the batch size.

In summary, the multi-objective version of Bayesian optimisation aims to iteratively approximate the true Pareto front, while the single-objective variant seeks to sequentially approach the global optimum. Both paradigms, model the objective(s) with a surrogate model(s) and employ an inexpensive acquisition functions to explore the design space for potentially optimal solutions.

2.2 Causality

Causality explores cause-and-effect relationships, focusing on dependencies between variables within potentially multivariate systems. It is a broad research domain with diverse subfields and an extensive body of literature. We review some basic concepts and state the rules of do-calculus, taken from Peters et al., 2017.

2.2.1 Structural causal models (SCMs)

In order to depict causal relationships and reason about causal effects, we require a formal representation of the underlying system. For this intent, so called *structural causal models* (Pearl, 2000) have become standard in the field of causality. This subsection leads up to the description of structural causal models, starting with some introductory graph terminology, commonly used across literature. We closely follow Peters et al., 2017 for the following definitions.

Definition 2.4 (Graph and subgraph)**.** A *graph* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is defined by a finite vertex set $\mathbf{V}$ and an edge set $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$, containing ordered pairs of distinct vertices. The *subgraph* of $\mathcal{G}$ restricted to $\mathbf{V}' \subseteq \mathbf{V}$ is given by $\mathcal{G}[\mathbf{V}'] = (\mathbf{V}', \mathbf{E}[\mathbf{V}'])$, where $\mathbf{E}[\mathbf{V}'] = \{(i, j) \in \mathbf{E} \mid i, j \in \mathbf{V}'\}$.

Definition 2.5 (Directed edge and graph)**.** An edge $(i, j) \in \mathbf{E}$ is said to be *directed*, denoted as $i \to j$, if $(i, j) \in \mathbf{E}$ and $(j, i) \notin \mathbf{E}$. Moreover, $\mathcal{G}$ is called a *directed graph* if all of its edges are directed.

Definition 2.6 (Path and cycle)**.** A *path* between $i, j \in \mathbf{V}$ is a sequence of distinct vertices $i = i_0,\ i_1,\ \ldots,\ i_n = j$ such that two consecutive vertices are adjacent, i.e. $(i_{k-1}, i_k) \in \mathbf{E}$ or $(i_k, i_{k-1}) \in \mathbf{E}$ for all $k = 1, 2, \ldots, n$. A *directed path* from i to j is a path in which all the arrows are going "forward", that is, $(i_{k-1}, i_k) \in \mathbf{E}$ for all $k = 1, 2, \ldots, n$. Moreover, a *cycle* is a directed path where the first and last vertices coincide, i.e. $i_0 = i_n$. A *directed acyclic graph* (DAG) is a directed graph with no cycles.

In other words, a DAG exhibits no directed paths from a node to itself. This property is required in order to arrange vertices in a causal ordering, enforcing the notion that causes must occur before their effects (Peters et al., 2017). Henceforth, every graph will be assumed to be a DAG.

Definition 2.7 (Parent, ancestor and descendant)**.** A vertex i is called a *parent* of j if $i \to j$.

Moreover, i is an *ancestor* of j if there exists a directed path from i to j; conversely, j is called a *descendant* of i.

For a vertex $i \in \mathbf{V}$, we denote the set if its parents, ancestors and descendants in $\mathcal{G}$ as $\mathrm{pa}(i)_\mathcal{G}$, $\mathrm{an}(i)_\mathcal{G}$ and $\mathrm{de}(i)_\mathcal{G}$, respectively. Here, it is assumed, that no vertex is a parent, an ancestor or a descendant of itself. Conversely, with a capital letter, we extend this notation to indeed include the argument in the result, e.g. $\mathrm{Pa}(i)_\mathcal{G} = \mathrm{pa}(i)_\mathcal{G} \cup \{i\}$. Moreover, we define these relations for sets of variables $\mathbf{V}' \subseteq \mathbf{V}$, e.g. $\mathrm{pa}(\mathbf{V}')_\mathcal{G} = \bigcup_{i \in \mathbf{V}'} \mathrm{pa}(i)_\mathcal{G}$ and $\mathrm{Pa}(\mathbf{V}')_\mathcal{G} = \bigcup_{i \in \mathbf{V}'} \mathrm{Pa}(i)_\mathcal{G}$. Equivalent conventions apply to the ancestor and descendant relationships.

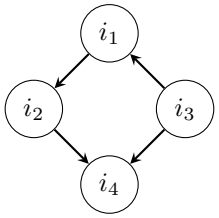
Definition 2.8 (d-separation)**.** Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG. A path π between vertices i_1 and i_n is said to be *blocked* by a set of vertices $\mathbf{S}$ (with neither i_1 nor i_n in $\mathbf{S}$) whenever there exists $i_k \in \pi$ such one of the following two possibilities occurs:

- It holds $i_k \in \mathbf{S}$, and directed edges in π do not meet head-to-head at i_k, i.e. either $i_{k-1} \to i_k \to i_{k+1}$, $i_{k-1} \leftarrow i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$;
- Neither i_k nor any of its descendants is in $\mathbf{S}$, and $i_{k-1} \to i_k \leftarrow i_{k+1}$.

Two disjoint subsets of vertices $\mathbf{A}$ and $\mathbf{B}$ are said to be *d-separated* by $\mathbf{S}$ if all paths between vertices in $\mathbf{A}$ and $\mathbf{B}$ are blocked by $\mathbf{S}$. Then, we write $\mathbf{A} \perp\!\!\!\perp_\mathcal{G} \mathbf{B} \mid \mathbf{S}$.

The concept of d-separation is indispensable for reasoning about causal effects, since it induces a notion of conditional independence (Peters et al., 2017) within variables of structural causal models (seen later), required for do-calculus. In order to illustrate this concept with an example, consider the DAG presented in Figure 2.1a. We observe that $\{i_2\}$ and $\{i_3\}$ are d-separated because all paths between them are blocked. More specifically, $\pi_1 : i_2 \leftarrow i_1 \leftarrow i_3$ is blocked by $\{i_1\}$, and $\pi_2 : i_2 \to i_4 \leftarrow i_3$ is blocked by $\{i_1\}$. With a similar argument, we infer that $\{i_1\}$ and $\{i_4\}$ are also d-separated.

Notably, DAGs are formally given as $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and are able to represent causal relationships in a binary manner. More specifically, a directed path between two variables indicates a causal relationship, while the absence of such signifies no causal connection. However, the quantitative



(a) A DAG with four nodes and four edges between them.

$$V_1 := f_{V_1}(V_3, U_{V_1}) = V_3 \cdot U_{V_1}$$
$$V_2 := f_{V_2}(V_1, U_{V_2}) = V_1 + U_{V_2}$$
$$V_3 := f_{V_3}(U_{V_3}) = U_{V_3}^2$$
$$V_4 := f_{V_4}(V_2, V_3, U_{V_4}) = (V_2 + V_3) \cdot U_{V_4}$$

(b) An SCM consisting of four structural assignments.

Figure 2.1: A DAG and associated structural assignments.

or functional aspects of these relationships cannot be captured with $\mathcal{G}$ alone. *Structural causal models* (SCMs) (Pearl, 2000) bridge this limitation and incorporate a description of the causal relations via functional assignments. For the formal definitions and results, we again closely follow Peters et al., 2017.

Definition 2.9 (Structural causal model (SCM))**.** A *structural causal model* $\mathcal{M} = \langle \mathbf{F}, P(\mathbf{U}) \rangle$, sometimes also called a *structural equation model*, consists of a collection $\mathbf{F} = \{f_1, \ldots, f_d\}$ of d *(structural) assignments* of the form

$$V_i := f_{V_i}(\mathrm{pa}(V_i), U_{V_i}), \tag{2.8}$$

where $\mathrm{pa}(V_i) \subseteq \{V_1, \ldots, V_d\} \backslash \{V_i\}$ is the set of parents of V_i; and $P(\mathbf{U})$ a joint distribution over the variables $\mathbf{U} = \{U_{V_1}, \ldots, U_{V_d}\}$ which are required to be jointly independent, that is, $P(\mathbf{U})$ is a product distribution. The graph $\mathcal{G}$ corresponding to $\mathcal{M}$ is constructed by creating a vertex for every V_i and placing directed edges from each parent in $\mathrm{pa}(V_i)$ to V_i.

Elements of $\mathbf{U}$ are also referred as *exogenous variables*, representing background factors outside the model, which nevertheless do affect the model. Conversely, variables in $\mathbf{V}$ are considered to be observable, and called *endogenous variables*. Henceforth, we use the notation $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ to refer to an SCM. Moreover, it is important to note that the structural assignments are to be interpreted as a set of assignments or functions, rather than a set of mathematical equations. For an example of an SCM, consider the structural assignments presented in Figure 2.1b and the associated graph given Figure 2.1b. In such as case, when the structural assignments in an SCM $\mathcal{M}$ are consistent with the causal structure encoded in a graph $\mathcal{G}$, we say that $\mathcal{M}$ *conforms* to $\mathcal{G}$.

Definition 2.9 indicates that there is a distinction between the following two SCMs

$$\mathbf{F}_1: \quad V_1 := U_{V_1},\ V_2 := 0 \cdot V_1 + U_{V_2}, \tag{2.9}$$

$$\mathbf{F}_2: \quad V_1 := U_{V_1},\ V_2 := U_{V_2}, \tag{2.10}$$

although it holds $0 \cdot V_1 = 0$. In order to omit such a distinction, we include the requirement that the assignments f_i must depend on all their input arguments along with the requirement that U_{V_i} takes non-zero values with a non-zero probability.

The following proposition formalises the probability distribution over the naturally occurring "states" of a given SCM.

Proposition 2.1. *An SCM* $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ *defines a unique distribution over* $\mathbf{V}$ *such that* $V_i \overset{d}{=} f_i(\mathrm{pa}(V_i)_{\mathcal{G}}, U_{V_i})$*,* $1 \leqslant i \leqslant d$*, called* observational distribution *and denoted* $P(\mathbf{V})$*.*

Here, $\overset{d}{=}$ denotes equality in probability distribution. For the proof of this statement, the reader is referred to Peters et al., 2017. It formalises the process of sampling n data points from the observational distribution $P(\mathbf{V})$: Firstly, we draw i.i.d. samples $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)} \sim P(\mathbf{U})$ from the

exogenous variables. Thereafter, we update the structural assignments (starting from source nodes and continuing with nodes of increasing number of parents) to obtain i.i.d. samples $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)} \sim P(\mathbf{V})$.

2.2.2 Interventions and confounders

Interventions are central to causality, enabling us to move beyond observations and actively manipulate and study causal systems. By intervening in a causal graph, one can isolate and quantify causal effects of specific variables. These manipulations are expressed using the do-operator (Pearl, 2000):

Definition 2.10 (Interventional distribution)**.** Let $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ be an SCM and $\mathcal{G}$ its associated graph. An *intervention* on variables $\mathbf{X} \subseteq \mathbf{V}$ involves replacing the structural equations f_X with a constant x, for all $X \in \mathbf{X}$. This action is denoted with the do-operator

$$\mathrm{do}(\mathbf{X} = \mathbf{x}), \tag{2.11}$$

where $\mathbf{x} = (x_1, \ldots, x_{|\mathbf{X}|})$. It removes the influence of all endogenous and exogenous variables on $X \in \mathbf{X}$. The graph $\mathcal{G}_{\overline{\mathbf{X}}}$ represents this intervention and is obtained by removing incoming edges into $\mathbf{X}$. The observational distribution of $\mathcal{G}_{\overline{\mathbf{X}}}$ is denoted

$$P(\mathbf{V}|\mathrm{do}(\mathbf{X} = \mathbf{X})) \tag{2.12}$$

and called *interventional distribution.*

For a variable $X \in \mathbf{X}$, we assume to have a feasible set of values to which it can be set to, referred to as its *interventional domain* and denoted by $\mathcal{D}(X)$. The *interventional domain* of a set of variables $\mathbf{X}$ is then defined as the Cartesian product $\mathcal{D}(\mathbf{X}) = \times_{X \in \mathbf{X}} \mathcal{D}(X)$. Notably, in the case $\mathbf{X} = \varnothing$, the do-operation does not perform any intervention, and the interventional and observational distributions coincide.

The interventional distribution describes the probabilistic behaviour of an SCM post intervention. Importantly, if can differ from the observational distribution. As an example, consider the SCM given by the structural assignments $V_1 = U_{V_1}$ and $V_2 = V_1 + U_{V_2}$, where $U_{V_1}, U_{V_2} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Then it holds $P(V_1|\mathrm{do}(V_2 = a)) \sim \mathcal{N}(0, 1)$ and $P(V_1|V_2 = a) \sim \mathcal{N}(a, 1)$.

Moreover, the observational distribution does not suffice for identifying and quantifying causal relationships between variables. It is necessary to consider the interventional distribution in order to infer how one variable affects another. This is especially relevant when the observational and interventional distributions differ. For this intent, we introduce the concept of *confounding* with the following definition from Peters et al., 2017.

Definition 2.11 (Confounding)**.** Let $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ be an SCM, where there is a directed path from X to Y, for $X, Y \in \mathbf{V}$. The causal effect from X to Y is called *confounded* if

$$P(Y|\mathrm{do}(X = x)) \neq P(Y|X = x), \tag{2.13}$$

that is, the interventional and observational distributions do not coincide.

Confounding occurs if there is a third variable Z influencing both X and Y (Peters et al., 2017). In this case, Z is called a *confounder*. It creates a spurious association between the variables and accounting for confounders is required for isolating causal effects. As an example, consider the SCM given in Figure 2.1, where the causal effect from V_1 to V_4 is confounded by V_3. A confounder is said to be *observed* if it is endogenous, and *unobserved* if it an exogenous variable. We adopt the notation $\mathbf{U}^X \subseteq \mathbf{U}$ for the set of exogenous variables affecting $X \in \mathbf{V}$, including U_X as well as unobserved confounders. The structural assignment is then given by $f_X(\text{pa}(X), \mathbf{U}^X)$.

2.2.3 Do-calculus

Do-calculus (Pearl, 1995) facilitates the identification of causal effects in non-parametric models, relying solely on the graph structure without restrictive assumptions on functional forms or probability distributions of an SCM. Specifically, do-calculus consists of three inference rules that permit the mapping of interventional and observational distributions under certain conditions dictated by the underlying causal graph. In this subsection, we state these three rules, but for a full derivation along with interpretations and connections to other topics in causality, the reader is referred to Pearl, 1995 and Peters et al., 2017.

Consider an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ and its associated DAG $\mathcal{G}$. Let $P(\mathbf{V})$ be the observational distribution. Moreover, $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ are disjoint subsets of endogenous variables. We denote by $\mathcal{G}_{\overline{\mathbf{X}}}$ the graph obtained by removing all incoming edges of $\mathbf{X}$ in $\mathcal{G}$. Similarly, $\mathcal{G}_{\underline{\mathbf{Z}}}$ refers to the graph obtained by removing all outgoing edges of $\mathbf{Z}$ in $\mathcal{G}$. To represent the deletion of both incoming and outgoing edges, we use the notation $\mathcal{G}_{\overline{\mathbf{X}},\underline{\mathbf{Z}}}$. Now, the *three rules of do-calculus* are given as follows (Peters et al., 2017):

1. **Insertion/deletion of observations**. An observation can be ignored if it is d-separated from the outcome. Formally,

$$P(\mathbf{Y}|\mathbf{Z}, \text{do}(\mathbf{X} = \mathbf{x}), \mathbf{W}) = P(\mathbf{Y}|\text{do}(\mathbf{X} = \mathbf{x}), \mathbf{W}) \quad \text{if } \mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{X}}}} \mathbf{Z} \mid \mathbf{W}, \mathbf{X}. \tag{2.14}$$

2. **Action/observation exchange**. Interventions and observations are interchangeable, when the causal effect of a variable on the outcome only influences the outcome through directed paths. Formally,

$$P(\mathbf{Y}|\text{do}(\mathbf{Z} = \mathbf{z}), \text{do}(\mathbf{X} = \mathbf{x}), \mathbf{W}) = P(\mathbf{Y}|\mathbf{Z}, \text{do}(\mathbf{X} = \mathbf{x}), \mathbf{W}) \quad \text{if } \mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{X}},\underline{\mathbf{Z}}}} \mathbf{Z} \mid \mathbf{W}, \mathbf{X}. \tag{2.15}$$

3. **Insertion/deletion of actions**. Interventions can be ignored if there are no unblocked causal paths to the outcome. Formally

$$P(\mathbf{Y}|\text{do}(\mathbf{Z} = \mathbf{Z}), \text{do}(\mathbf{X} = \mathbf{x}), \mathbf{W}) = P(\mathbf{Y}|\text{do}(\mathbf{X} = \mathbf{x}), \mathbf{W}) \quad \text{if } \mathbf{Y} \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathbf{X}},\overline{\mathbf{Z(W)}}}} \mathbf{Z} \mid \mathbf{W}, \mathbf{X}, \tag{2.16}$$

where $\mathbf{Z}(\mathbf{W}) \subseteq \mathbf{Z}$ is a subset of variables such that it does not contain ancestors of any variable in $\mathbf{W}$.

For the "completeness" proofs for the three rules of do-calculus, we refer to Huang and Valtorta, 2006 and Shpitser and Pearl, 2006. Their contributions include a complete characterisation of the conditions under which causal effects can be identified from the observational distribution.

In summary, we have firstly introduced structural causal models that provide a framework for understanding and quantifying causal relationships within a system. By formalising the relationships between endogenous and exogenous variables, they enable the study of observational and interventional distributions. The introduction of concepts such as confounding and the do-operator highlights the necessity of interventions for isolating causal effects. Furthermore, the rules of do-calculus offer a tool for identifying causal effects in non-parametric models, relying solely on the causal graph structure.

3 Related work

This chapter presents an overview of the research areas related to *multi-objective causal Bayesian optimisation* (MO-CBO), the novel problem formulation introduced in this thesis. The aim of MO-CBO is to sequentially optimise multiple target variables within a causal system, placing it at the intersection of multi-objective optimisation and causality theory. Moreover, this work contributes to the broader field of *causal decision-making*, extending the subfield of *causal Bayesian optimisation* to handle multi-objective problems. The research contribution is visually represented in Figure 3.1, highlighting the connection between causal decision-making, multi-objective optimisation, and causality. In the following sections, we provide a concise review of these research areas.

Recently, causal decision-making has been placed in the field of *causal AI*, which focuses on developing algorithms and systems that understand, reason about, and leverage cause-and-effect relationships to improve predictions, interventions, and decision-making.

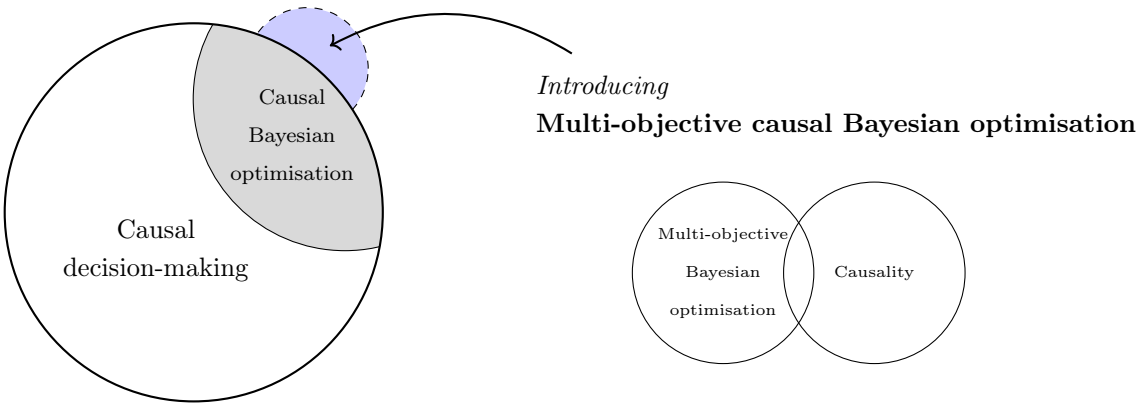


Figure 3.1: Research context of multi-objective causal Bayesian optimisation (MO-CBO).

Multi-objective Bayesian optimisation

Bayesian optimisation. Bayesian optimisation was initially formalised by Močkus, 1975, and later popularised with the *efficient global optimisation* algorithm (Jones et al., 1998). It is a sample-efficient paradigm for optimising expensive-to-evaluate and unknown objective functions. More specifically, a Bayesian optimisation algorithm relies on two critical components: a

surrogate model and an acquisition function. The surrogate model, typically a regression-based model such as a *Gaussian process* (Rasmussen, 2004), approximates the objective function and provides probabilistic predictions as well as uncertainty estimates. The acquisition function utilises these predictions for selecting the next evaluation point, achieving a balance between exploitation of regions with high predicted performance and exploration of areas with greater uncertainty. Popular acquisition functions include *probability of improvement* (Kushner, 1964), *expected Improvement* (Jones et al., 1998), and *upper confidence bound* (Srinivas et al., 2010). The interplay between the surrogate model and acquisition function makes Bayesian optimisation highly effective for problems where each function evaluation is expensive, such as hyperparameter tuning or design optimisation.

Multi-objective optimisation. Multi-objective optimisation problems involve several, potentially conflicting, objective functions that must be optimised simultaneously by identifying a set of Pareto-optimal solutions (Miettinen, 1999). For this intent, so called *population-based multi-objective evolutionary algorithms* have been developed, such as NSGA-II (Deb et al., 2002) and MOEA/D (Zhang & Li, 2007). These methods are highly popularised and widely used for multi-objective optimisation problems across various domains, such as manufacturing design (Schulz et al., 2018) and optimal control (Gambier & Badreddin, 2007). However, they generally require numerous function evaluations due to the inherent nature of evolutionary computation, making them impractical for many real-world applications.

Multi-objective Bayesian optimisation. Combining principles from multi-objective optimisation and Bayesian optimisation, multi-objective Bayesian optimisation is designed to solve problems with several expensive-to-evaluate objective functions (Laumanns & Ocenasek, 2002). These functions are typically modelled independently via Gaussian Processes. Advanced acquisition functions have been developed here, including *expected hypervolume improvement* (Couckuyt et al., 2014) and *predictive entropy search* (Henrández-Lobato et al., 2014). Furthermore, there are two main selection strategies: Single-point methods select and evaluate one candidate solution at each iteration, while batch methods select multiple solutions simultaneously for parallel evaluation.

One of the earliest single-point algorithms is ParEGO (Knowles, 2006), which randomly scalarises the multi-objective problem into a single-objective problem and chooses a sample that maximises the expected improvement. In another contribution, PAL (Zuluaga et al., 2013) aims to improve performance by reducing the uncertainty on the surrogate model. USeMO (Belakaria et al., 2020) employs maximum uncertainty as a selection criterion, based on the set of Pareto-optimal points produced by NSGA-II, and demonstrates state-of-the-art performance.

Regarding batch methods, MOEA/D-EGO (Zhang et al., 2010) is recognised as one of the earliest approaches, extending ParEGO to incorporate multiple scalarisation weights and perform parallel optimisation through MOEA/D. In another line of work, TSEMO (Bradford et al.,

2018) adopts Thompson Sampling on the Gaussian Process posterior as an acquisition function, optimises multiple objectives with NSGA-II, and selects the next batch of samples by maximising the hypervolume improvement. Moreover, DGEMO (Konakovic Lukovic et al., 2020) proposes a novel acquisition method that considers diversity in both input and performance space. Finally, qNEHVI (Daulton et al., 2021) is based on expected hypervolume improvement and scales to highly parallel evaluations of noisy objectives.

Causality

Causality is the study of cause-and-effect relationships, aiming to understand dependencies between variables in a possibly multivariate system. With a vast and diverse body of literature (see for example Peters et al., 2017 and Pearl, 2009), it is a broad research field that encompasses various sub-branches. Among these, *causal discovery* (Glymour et al., 2019) focuses on identifying causal relationships from data, while *causal inference* (Yao et al., 2021) seeks to determine causal effects, particularly under interventions. In this thesis, the focus is set on causal inference. The formal framework for causal inference is largely attributed to Pearl, 1995 with the derivation of *do-calculus*, an comprehensive mathematical system for inferring causal effects from observations. Central to this system are so called *structural causal models* that represent a causal system with a directed acyclic graph and deterministic, noise-induced functional assignments between the vertices.

The study of causality has been proposed as a mean to overcome the limitations of traditional regression methods that capture correlation instead of causation, critiquing established fields such as machine learning, deep learning, and Bayesian optimisation (Pearl & Mackenzie, 2018). Incorporating causality into these areas has led to the emergence of *causal AI* and new paradigms such as *causal machine learning* (Kaddour et al., 2022), *causal deep learning* (Berrevoets et al., 2023), and *causal Bayesian optimization* (Aglietti et al., 2020). Causal AI is a research field focused on developing causally-informed autonomous systems capable of reasoning, interpreting, and explaining their decision-making process. Xia et al., 2021 show that while neural networks possess universal approximability, they cannot learn structural causal models or predict intervention effects from observational data alone. Interestingly, Richens and Everitt, 2024 demonstrate that any agent capable of satisfying a regret bound under a large set of distributional shifts must have learned an approximate causal model. Although some of the work in this field focuses on the inherent properties of current machine learning systems, causal AI intersects with *causal decision-making*, encompassing the literature discussed in the next section.

Causal decision-making

Studying causality has been shown to be of critical importance for decision-making in various practical domains such as engineering, healthcare and economics (Pearl & Mackenzie, 2018). Leveraging tools from causal inference to make causally-informed decisions is a field called

causal decision-making. Albeit still limited, the literature on causal decision making has recently been perceived with rising significance.

There is a line of work in causal-decision making focusing on *multi-armed bandit* problems and *reinforcement learning* settings. Here, actions or arms correspond to interventions on an arbitrary causal graph with existing links between the agent's decisions and the received rewards. Bareinboim et al., 2015 consider multi-armed bandit problems where unobserved confounders affect arms as well as rewards, and demonstrate that both experimental and observational data is necessary for minimising regret. Lu et al., 2018 focus on reinforcement learning settings with unobserved confounders, and extend the Actor-Critic method to its deconfounding variant. Lee and Bareinboim, 2018 identify a set of possibly-optimal arms that an agent should play in order to maximise its expected reward in a multi-armed bandit problem. Moreover, Lee and Bareinboim, 2019 extend their previous work to graphs with non-manipulable variables. Lattimore et al., 2016 study a specific family of multi-armed bandit problems called parallel bandits in which n factors affect the reward independently and there are $2n$ possible interventions. In another line of work, Buesing et al., 2019 utilise structural causal models to perform counterfactual evaluations of arbitrary policies based on individual off-policy episodes. Foerster et al., 2018 address the multi-agent setting by introducing a framework where each agent learns using a shaped reward. This reward is derived by comparing the global reward to the counterfactual reward obtained when the agent's action is substituted with a default action.

Causal Bayesian optimisation. Aglietti et al., 2020 introduce *causal Bayesian optimisation* (CBO), an algorithm which generalises Bayesian optimisation to scenarios where causal relationships between variables are known (while the functional relations between variables remain unknown). There is a growing body of research specifically focused on advancing CBO. Aglietti et al., 2021 propose *dynamic causal Bayesian optimisation*, and explore the problem of performing a sequence of optimal interventions in causal dynamical systems where both the outcome variable and the inputs evolve over time. In another contribution, Aglietti et al., 2023 introduce *constrained causal Bayesian optimisation*, a framework that enforces constraints on certain variables throughout the sequential optimisation process, even in the absence of direct interventions on those variables. Moreover, Branchini et al., 2023 derive *causal entropy optimisation*, a CBO variant that addresses causal structure uncertainty with an information-theoretic acquisition function, enabling simultaneous causal graph learning and causal effect optimisation. Gultchin et al., 2023 propose *functional causal Bayesain optimisation*, allowing functional interventions where a variable is set as a deterministic function of other variables in the causal graph. Sussex et al., 2023 propose *mode-based causal Bayesian optimisation* which learns full causal mechanisms and bounds regret non-asymptotically. Furthermore, Sussex et al., 2024 introduce a framework called *adversarial causal Bayesian optimisation*, in which other agents or external events can manipulate the system and perform adversarial interventions. Ren and Qian, 2024 extend CBO by learning the distribution of exogenous variables, allowing the incorporation of causal models beyond additive noise models. Zeitler and Astudillo, 2024 consider extending *causal entropy search* by incorporating preference elicitation, a technique

used in Bayesian optimisation to gather knowledge from experts. Anonymous, 2024 (under blind review at ICLR 2025) view the CBO problem under the cumulative regret objective with unknown causal graphs, considering both soft and hard interventions.

To the best of our knowledge, a multi-objective variant of CBO has not yet been proposed and shall be the contribution of this thesis.

4 Multi-objective SCMs: where to intervene?

Sequential decision-making on causal systems is an optimisation paradigm that is central to this thesis as we aim to extend *causal Bayesian optimisation* (Aglietti et al., 2020) to the multi-target setting. It involves sequentially implementing interventions and manipulating variables in a causal system to achieve desired outcomes. One strategy is to intervene on all variables simultaneously, disregarding the underlying causal relationships. However, empirical evidence demonstrates that this approach can lead to substantial regret (Lee & Bareinboim, 2018). Another naive strategy is to alternately explore all intervention sets. Albeit, such a brute-force exploration is indeed a valid approach, its complexity rises exponentially as the number of treatment variables increases. These observations necessitate more efficient strategies. Lee and Bareinboim, 2018 address this problem in the multi-armed bandit setting by leveraging the graph topology to identify intervention sets that are redundant to consider. Their formalism exploits the rules of do-calculus to identify invariances and partial-orders among intervention sets, to obtain those sets that could potentially yield optimal outcomes for a given graph. Both scenarios, with and without unobserved confounders, are considered. However, their formalism is limited to scenarios where a single target variable is to be optimised. To take advantage of their ideas for this thesis, the relevant concepts and their theoretical properties must be extended to accommodate multi-target settings, which will be the focus of this chapter.

4.1 Problem formulation

Notation. Let $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ be a structural causal model (SCM) and $\mathcal{G}$ its associated graph that encodes the underlying causal mechanisms. Recall that $\mathbf{U}$ is a set of independent exogenous variables distributed according to the probability distribution $P(\mathbf{U})$, $\mathbf{V}$ is a set of endogenous variables, and $\mathbf{F} = \{f_V\}_{V \in \mathbf{V}}$ is a set of functions such that $V = f_V(\text{pa}(V)_{\mathcal{G}}, \mathbf{U}^V)$. Within $\mathbf{V}$ there are two different types of variables to be distinguished: treatment variables $\mathbf{X}$, that can be set to specific values, and output variables $\mathbf{Y} = \{Y_1, \dots, Y_m\}$, that represent the outcomes of interest and are non-manipulative. A set $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ is called an *intervention set* and $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ an *intervention value*. The graph that is obtained after removing incoming edges into $\mathbf{X}_s$ is denoted $\mathcal{G}_{\overline{\mathbf{X}}_s}$ and represents the intervention on $\mathbf{X}_s$. Given two intervention sets $\mathbf{X}_s, \mathbf{X}'_s$ and $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$, we denote by $\mathbf{x}_s[\mathbf{X}'_s]$ the values of $\mathbf{x}_s$ corresponding to $\mathbf{X}_s \cap \mathbf{X}'_s$. Moreover, $\text{CC}(\mathbf{X}_s)_{\mathcal{G}}$ denotes a c-component of $\mathcal{G}$ (Tian & Pearl, 2002), which, in this context, is the maximal set of variables that includes $\mathbf{X}_s$ and is connected via unobserved confounders. For more details and notation on these concepts from causal inference, the reader is referred to section 2.2.

Assumptions. In this chapter, we assume that it holds $\mathbf{Y} \cap \text{an}(\mathbf{Y})_{\mathcal{G}} = \varnothing$, that is, no target

variable $Y_i \in \mathbf{Y}$ is parent to another $Y_j \in \mathbf{Y}$. Additionally, $\mathbf{U}^{Y_i} \cap \mathbf{U}^{Y_j} = \varnothing$ shall ensure that there is no unobserved confounder between any two target variables. Moreover, we assume that the causal relationships encoded in $\mathcal{G}$ are fully accessible while the underlying parametrisation, i.e. $\mathbf{F}$ and $P(\mathbf{U})$, remains unknown. This restricted information is denoted as $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$. The assumption allows generalisation across systems that share the same causal structure.

Problem statement

In the multi-objective setting, the aim is to identify intervention sets that can minimise (or, for the dual problem, maximise) all target variables $Y_1, \ldots, Y_m$ simultaneously. The outcomes of an intervention $\text{do}(\mathbf{X}_s = \mathbf{x}_s)$ are captured as the expected values of each target:

$$\mathbb{E}_{P(Y_i|\text{do}(\mathbf{X}_s=\mathbf{x}_s))}[Y_i] := \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)], \tag{4.1}$$

where $P(Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s))$ denotes the interventional distribution of Y_i, for all $i = 1, \ldots, m$. The goal is to identify intervention sets that can establish non-dominated trade-offs between the objectives $\mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$, $1 \leqslant i \leqslant m$. We aim to develop a methodology that is agnostic to the underlying parametrisation of a given causal graph. The following definition of optimality on $\mathcal{P}(\mathbf{X})$ satisfies this requirement:

Definition 4.1 (Possibly-optimal intervention set)**.** Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ is called *possibly-optimal* if there exists $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ such that for all other intervention sets $\mathbf{X}'_s \in \mathcal{P}(\mathbf{X})$ and $\mathbf{x}'_s \in \mathcal{D}(\mathbf{X}_s)$ it holds $\mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)] \geqslant \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$ for all $1 \leqslant i \leqslant m$, and $\mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)] > \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$ for at least one $1 \leqslant i \leqslant m$, for some SCM conforming to $\mathcal{G}$

In other words, an intervention set is possibly-optimal, if there exists an intervention value such that it establishes a dominant solution across the averaged targets. We use the term *possibly-optimal*, because an intervention set might be optimal for one SCM, conforming to $\mathcal{G}$, but not necessarily for others. Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, we define the problem objective of this chapter as identifying all intervention sets that are possibly-optimal, formally,

$$\{\mathbf{X}_s \in \mathcal{P}(\mathbf{X}_s) \mid \mathbf{X}_s \text{ is possibly-optimal}\}. \tag{4.2}$$

The challenge here lies in leveraging the known graph structure to identify such intervention sets.

4.2 Equivalence among intervention sets

As a first step, Lee and Bareinboim, 2018 establish invariances within $\mathcal{P}(\mathbf{X})$ in regards to the effects of intervention sets on the target variables. With the following definition and proposition, we extend their notions to the multi-objective setting.

Definition 4.2 (Minimal intervention set)**.** Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, a set $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ is called a *minimal intervention set* if there is no subset $\mathbf{X}'_s \subset \mathbf{X}_s$ such that for each $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ it holds $\mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)] = \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}_s[\mathbf{X}'_s])]$, $1 \leqslant i \leqslant m$, for every SCM conforming to $\mathcal{G}$.

We denote the set of minimal intervention sets with $\mathbb{M}_{\mathcal{G},\mathbf{Y}}$. In other words, no subset of a minimal intervention set can achieve the same expected outcome on $\mathbf{Y}$. As such, intervention sets, that are not *minimal* in the sense of Definition 4.2, are redundant to consider in any sequential decision-making paradigm that aims to minimise multiple target variables simultaneously. The next proposition addresses the question of identifying minimal intervention sets in a given causal graph with multiple target variables.

Proposition 4.1. *Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ is a minimal intervention set if and only if it holds $\mathbf{X}_s \subseteq \text{an}(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}_s}}$.*

Proof. (If) Let $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ be any intervention value. Assume that there is a subset $\mathbf{X}'_s \subset \mathbf{X}_s$ such that $\mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)] = \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}_s[\mathbf{X}'_s])]$ for all $1 \leqslant i \leqslant m$. Consider an SCM with real-valued variables where each $V \in \mathbf{V}$ is associated with its own binary exogenous variable U_V with $P(U_V = 1) = 0.5$. Let the function of an endogenous variable be the sum of values of its parents. For the sake of contradiction, assume $\mathbf{X}_s \subseteq \text{an}(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}_s}}$. Then, there exists a directed path from $\mathbf{X}_s \backslash \mathbf{X}'_s$ to some Y_i without passing $\mathbf{X}'_s$. Hence, setting $\mathbf{W} = \mathbf{X}_s \backslash \mathbf{X}'_s$ to the values $\mathbf{w} = \mathbb{E}[\mathbf{W}|\text{do}(\mathbf{X}'_s = \mathbf{x}_s[\mathbf{X}'_s])] + 1$ yields $\mathbb{E}[Y_i|\text{do}(\mathbf{W} = \mathbf{w}, \mathbf{X}'_s = \mathbf{x}_s[\mathbf{X}'_s])] > \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}_s[\mathbf{X}'_s])]$, contradicting the assumption.

(Only if) Assume that $\mathbf{X}_s \nsubseteq \text{an}(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}_s}}$. Then, for $\mathbf{X}'_s = \mathbf{X}_s \cap \text{an}(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}_s}}$ it holds $\mathbf{X}'_s \subset \mathbf{X}_s$ and by the third rule of do-calculus, for every $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ it holds $\mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)] = \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}_s[\mathbf{X}'_s])]$, $1 \leqslant i \leqslant m$. This is a contradiction because $\mathbf{X}_s$ was assumed to be a minimal intervention set. □

We conclude that Proposition 4.1 provides a complete characterisation of the minimal intervention sets in $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$. A visual depiction of the result is presented in Figure 4.1 contrasting an intervention set that is minimal, as seen in Figure 4.1a, to one that is not, see Figure 4.1b. In the latter case, we observe that some variables in $\mathbf{X}_s$ do not propagate their effects to any Y_i since $\mathbf{X}_s \nsubseteq \text{an}(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}_s}}$. By removing such variables, $\mathbf{X}_s$ becomes a minimal intervention set.

(a) $\mathbf{X}_s$ is a minimal intervention set (b) $\mathbf{X}_s$ is not a a minimal intervention set

Intervened variable in $\mathbf{X}_s$ Target variable

Figure 4.1: Graphical characterisation of minimal intervention sets $\mathbf{X}_s$, comparing a minimal intervention set to one that is not minimal.

4.3 Partial-orders among intervention sets

Lee and Bareinboim, 2018 introduce partial-orders among minimal intervention sets that can disregard some of those sets as inefficient. In this section, we formalise the respective multi-objective version with regard to the problem statement from section 4.1: Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, the question arises whether there exists an SCM, conforming to $\mathcal{G}$, such that a given intervention set $\mathbf{X}_s$ is possibly-optimal. In this section, we derive graphical criteria to identify all intervention sets that are indeed possibly-optimal for some SCM, conforming to $\mathcal{G}$. All definitions and results leading up to Theorem 4.1 are a multi-objective extension of the work conducted by Lee and Bareinboim, 2018. Firstly, we integrate the concept of minimal intervention sets with that of possibly-optimal sets.

Definition 4.3 (Possibly-optimal minimal intervention set)**.** Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, let $\mathbf{X}_s$ be a minimal intervention set. Then, $\mathbf{X}_s$ is called *possibly-optimal* if there exists $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ such that for all $\mathbf{X}'_s \in \mathbb{M}_{\mathcal{G},\mathbf{Y}} \backslash \mathbf{X}_s$ and $\mathbf{x}'_s \in \mathcal{D}(\mathbf{X}_s)$ it holds $\mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)] \geqslant \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$ for all $1 \leqslant i \leqslant m$, and $\mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)] > \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$ for at least one $1 \leqslant i \leqslant m$, for some SCM conforming to $\mathcal{G}$

We denote the set of possibly-optimal minimal intervention sets with $\mathbb{P}_{\mathcal{G},\mathbf{Y}}$. Identifying possibly-optimal minimal intervention sets is a non-trivial task, especially in the presence of unobserved confounders. For simplicity, we first consider the special case in which $\mathcal{G}$ exhibits no unobserved confounders between Y_i and any of its ancestors. The following proposition fully characterises possibly-optimal minimal intervention sets in such a scenario.

Proposition 4.2. *Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, if no Y_i is confounded with $\text{an}(Y_i)_\mathcal{G}$ via unobserved confounders, then $\text{pa}(\mathbf{Y})_\mathcal{G}$ is the only possibly-optimal minimal intervention set.*

Proof. Let $\mathbf{X}_s = \text{pa}(\mathbf{Y})_\mathcal{G}$, and let $\text{pa}(\mathbf{Y})_\mathcal{G} \neq \mathbf{X}'_s$ be another minimal intervention set. Assume for contradiction that $\mathbf{X}'_s$ is a possibly-optimal minimal intervention set. Then, there exists $\mathbf{x}'_s \in \mathcal{D}(\mathbf{X}'_s)$ such that for all $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ it holds $\mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)] \geqslant \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)]$ for all $1 \leqslant i \leqslant m$, and $\mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)] > \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)]$ for at least one $1 \leqslant i \leqslant m$. Moreover, define $\mathbf{Z} = \mathbf{X}_s \backslash (\mathbf{X}'_s \cap \mathbf{X}_s)$. Then, it holds

$$\mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)] = \int_{\mathcal{D}(\mathbf{Z})} \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s), \mathbf{Z} = \mathbf{z}] P(\mathbf{Z} = \mathbf{z}|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)) d\mathbf{z} \tag{4.3}$$

$$= \int_{\mathcal{D}(\mathbf{Z})} \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s, \mathbf{Z} = \mathbf{z})] P(\mathbf{Z} = \mathbf{z}|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)) d\mathbf{z} \tag{4.4}$$

$$= \int_{\mathcal{D}(\mathbf{Z})} \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s \cap \mathbf{X}_s = \mathbf{x}'_s[\mathbf{X}_s], \mathbf{Z} = \mathbf{z})] P(\mathbf{Z} = \mathbf{z}|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)) d\mathbf{z} \tag{4.5}$$

for all $i = 1, \ldots, m$. Note that the second and third equalities are derived through the second and third rules of do-calculus, respectively. Moreover, in the third equality, we also assume that Y_i is not confounded with $\text{an}(Y_i)_\mathcal{G}$ via unobserved confounders. We define $\mathbf{x}_s(\mathbf{z}) \in \mathcal{D}(\mathbf{X}_s)$

such that $\mathbf{x}_s(\mathbf{z})[\mathbf{X}'_s] = \mathbf{x}'_s[\mathbf{X}_s]$ and $\mathbf{x}_s(\mathbf{z})[\mathbf{Z}] = \mathbf{z}$. Then, for some $1 \leqslant i \leqslant m$ it holds

$$\mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)] = \int_{\mathcal{D}(\mathbf{Z})} \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s(\mathbf{z})]P(\mathbf{Z} = \mathbf{z}|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s))d\mathbf{z} \tag{4.6}$$

$$> \int_{\mathcal{D}(\mathbf{Z})} \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)]P(\mathbf{Z} = \mathbf{z}|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s))d\mathbf{z} \tag{4.7}$$

$$= \mathbb{E}[Y_i|\text{do}(\mathbf{X}'_s = \mathbf{x}'_s)], \tag{4.8}$$

where the inequality is a result of the assumption that $\mathbf{X}'_s$ is a possibly-optimal minimal intervention set. This is a contradiction, and we conclude that $\mathbf{X}'_s$ cannot be possibly-optimal. However, for $\mathbf{X}_s = \text{pa}(\mathbf{Y})_\mathcal{G}$, it is possible to construct an SCM, conforming to $\mathcal{G}$, such that it it yields a dominant solution (see for example the proof of Proposition 4.4). This shows that $\text{pa}(\mathbf{Y})_\mathcal{G}$ can be the only possibly-optimal minimal intervention set. □

We now continue and study possibly-optimal minimal intervention sets in the more general case where unobserved confounders can be present between Y_i and any of its ancestors. For this intent, we extend two existing concepts, called *minimal unobserved-confounders' territory* and *interventional border* (Lee & Bareinboim, 2018), to the multi-objective setting. Using these notions, we derive results which can fully characterise possibly-optimal minimal intervention sets in the aforementioned scenario.

Definition 4.4 (Minimal unobserved confounders' territory)**.** Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, let $\mathcal{H}$ be defined as the subgraph $\mathcal{G}[\text{An}(\mathbf{Y})_\mathcal{G}]$. A set of variables $\mathbf{T}$ in $\mathcal{H}$, containing $\mathbf{Y}$, is called a *UC-territory* for $\mathcal{G}$ w.r.t. $\mathbf{Y}$ if $\text{De}(\mathbf{T})_\mathcal{H} = \mathbf{T}$ and $\text{CC}(\mathbf{T})_\mathcal{H} = \mathbf{T}$. The UC-territory $\mathbf{T}$ is said to be *minimal*, denoted $\mathbf{T} = \text{MUCT}(\mathcal{G}, \mathbf{Y})$, if no $\mathbf{T}' \subset \mathbf{T}$ is a UC-territory.

A minimal UC-territory for $\mathcal{G}$ w.r.t. $\mathbf{Y}$ can be constructed by extending a set of variables, starting from $\mathbf{Y}$, and iteratively updating the set with the c-component and descendants of the set. More intuitively, it is the minimal subset of $\text{An}(\mathbf{Y})_\mathcal{G}$ that is governed by unobserved confounders, where at least one target Y_i is adjacent to an unobserved confounder.

Definition 4.5 (Interventional border)**.** Let $\mathbf{T}$ be a minimal UC-territory for $\mathcal{G}$ w.r.t. $\mathbf{Y}$. Then, $\mathbf{B} = \text{pa}(\mathbf{T})_\mathcal{G} \backslash (\text{pa}(\mathbf{T})_\mathcal{G} \cap \mathbf{T})$ is called *interventional border* for $\mathcal{G}$ w.r.t. $\mathbf{Y}$, denoted $\text{IB}(\mathcal{G}, \mathbf{Y})$.

We illustrate these two concepts in Figure 4.2a where the minimal UC-territory and interventional border are depicted with pink and blue nodes, respectively. Starting from $\mathbf{T} = \{Y_1, Y_2\}$, the minimal UC-territory is constructed as follows: Since X_3 has an unobserved confounder with Y_1, we update $\mathbf{T} = \{Y_1, Y_2, X_3\}$, and thereafter add all descendants of X_3 obtaining $\mathbf{T} = \{Y_1, Y_2, X_3, X_1\}$. Now, $X_1 \in \mathbf{T}$ has an unobserved confounder with X_2, so we again add X_2 and its descendants, and update to $\mathbf{T} = \{Y_1, Y_2, X_3, X_1, X_2\}$. Since there are no more unobserved confounders between $\mathbf{T}$ and $\text{An}(\mathbf{Y})_\mathcal{G} \backslash \mathbf{T}$, the minimal UC-territory has been found and is given by $\text{MUCT}(\mathcal{G}, \mathbf{Y}) = \{Y_1, Y_2, X_3, X_1, X_2\}$. The interventional border is, by definition, $\text{IB}(\mathcal{G}, \mathbf{Y}) = \{X_4, X_5\}$. Before connecting the notion of minimal UC-territory and interventional

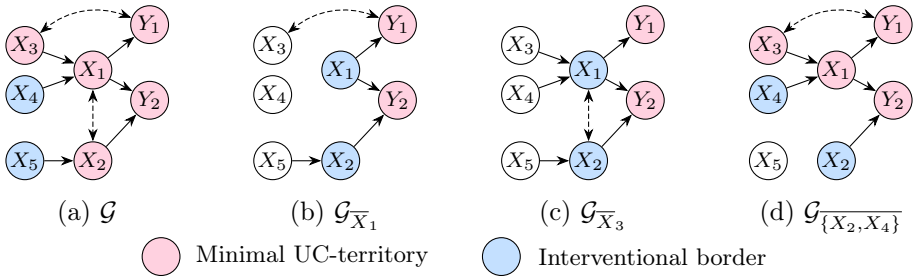


Figure 4.2: Depiction of the minimal UC-territory and interventional border in a causal graph $\mathcal{G}$ and several of its intervened graphs. Dashed bi-directed edges represent unobserved confounders.

border to possibly-optimal minimal intervention sets, we extend an important proposition from Lee and Bareinboim, 2018 to the multi-objective setting.

Proposition 4.3. *Let* $\mathbf{T}$ *be a minimal UC-territory and* $\mathbf{B}$ *an interventional border for* $\mathcal{G}$ *w.r.t.* $\mathbf{Y}$*. Let* $\mathbf{X}_s \subseteq \mathbf{X}$ *be an intervention set and* $\mathbf{S} = (\mathbf{T} \cap \mathbf{X}_s) \cup \mathbf{B}$*. Then, there exists no* $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ *such that for all* $\mathbf{s} \in \mathcal{D}(\mathbf{S})$ *it holds* $\mathbb{E}[Y_i|\mathrm{do}(\mathbf{S} = \mathbf{s})] \geqslant \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)]$ *for all* $1 \leqslant i \leqslant m$*, and* $\mathbb{E}[Y_i|\mathrm{do}(\mathbf{S} = \mathbf{s})] > \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)]$ *for some* $1 \leqslant i \leqslant m$*.*

Proof. (Case $\mathbf{B} \subseteq \mathbf{X}_s$) Let $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ be an intervention value. Then, by the third rule of do-calculus, it holds $\mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)] = \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s \cap (\mathbf{T} \cup \mathbf{B}) = \mathbf{x}_s[\mathbf{T} \cup \mathbf{B}])]$, $1 \leqslant i \leqslant m$. Since $\mathbf{X}_s \cap (\mathbf{T} \cup \mathbf{B}) = \mathbf{S}$, by setting $\mathbf{s} = \mathbf{x}_s[\mathbf{T} \cup \mathbf{B}]$, it follows $\mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)] = \mathbb{E}[Y_i|\mathrm{do}(\mathbf{S} = \mathbf{s})]$.

(Case $\mathbf{B} \nsubseteq \mathbf{X}_s$) Assume for contradiction that there exists $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ such that for all $\mathbf{s} \in \mathcal{D}(\mathbf{S})$ it holds $\mathbb{E}[Y_i|\mathrm{do}(\mathbf{S} = \mathbf{s})] \geqslant \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)]$ for all $1 \leqslant i \leqslant m$, as well as $\mathbb{E}[Y_i|\mathrm{do}(\mathbf{S} = \mathbf{s})] > \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)]$ for some $1 \leqslant i \leqslant m$. Let $\mathbf{B}' = \mathbf{B} \backslash (\mathbf{X}_s \cap \mathbf{B})$. Then, it holds

$$\mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)] = \int_{\mathcal{D}(\mathbf{B}')} \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s), \mathbf{B}' = \mathbf{b}'] P(\mathbf{B}' = \mathbf{b}'|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)) d\mathbf{b}' \tag{4.9}$$

for all $i = 1, \ldots, m$. The first term becomes

$$\begin{aligned} \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s), \mathbf{B}' = \mathbf{b}'] &= \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s, \mathbf{B}' = \mathbf{b}')] && (4.10) \\ &= \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s \cap (\mathbf{T} \cup \mathbf{B}) = \mathbf{x}_s[\mathbf{T} \cup \mathbf{B}], \mathbf{B}' = \mathbf{b}')] && (4.11) \end{aligned}$$

where the first and second equalities result from the second and third rules of do-calculus, respectively. Since $(\mathbf{X}_s \cap (\mathbf{T} \cup \mathbf{B})) \cup \mathbf{B}' = \mathbf{S}$, define $\mathbf{s}(\mathbf{b}') \in \mathcal{D}(\mathbf{S})$ such that $\mathbf{s}(\mathbf{b}')[\mathbf{X}_s \cap (\mathbf{T} \cup \mathbf{B})] = \mathbf{x}_s[\mathbf{T} \cup \mathbf{B}]$ and $\mathbf{s}(\mathbf{b}')[\mathbf{B}'] = \mathbf{b}'$. Then, for some $1 \leqslant i \leqslant m$ it holds

$$\begin{aligned} \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)] &= \int_{\mathcal{D}(\mathbf{B}')} \mathbb{E}[Y_i|\mathrm{do}(\mathbf{S} = \mathbf{s}(\mathbf{b}'))] P(\mathbf{B}' = \mathbf{b}'|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)) d\mathbf{b}' && (4.12) \\ &> \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)]. && (4.13) \end{aligned}$$

This is a contradiction, which proves the claim of this proposition. □

Notably, Proposition 4.3 indicates that every possibly-optimal minimal intervention set has to contain the interventional border and does not intersect with the ancestors of the interventional border. For instance, the intervention set $\mathbf{X}_s = \{X_2, X_3, X_4\}$ in Figure 4.3, is not possibly-optimal as it does not contain the entire interventional border, i.e. $\{X_4, X_5\} \nsubseteq \mathbf{X}_s$. By Proposition 4.3, intervening on $\mathbf{S} = \{X_2, X_3, X_4, X_5\}$ is at least as optimal as intervening on $\mathbf{X}_s$. The following proposition, which we extended from (Lee & Bareinboim, 2018) for the multi-target setting, is a building block for characterising possibly-optimal minimal intervention sets via interventional borders.

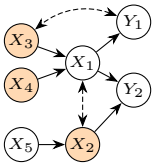


Figure 4.3: Causal graph $\mathcal{G}$ with orange nodes representing an intervention set $\mathbf{X}_s$.

Proposition 4.4. *Given* $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, $\mathrm{IB}(\mathcal{G}, \mathbf{Y})$ *is a possibly-optimal minimal intervention set.*

Proof. The intuition of this proof is to construct an SCM, conforming to $\mathcal{G}$, for which the single best strategy involves intervening on $\mathrm{IB}(\mathcal{G}, \mathbf{Y})$. Let $\mathbf{T}$ and $\mathbf{B}$ denote $\mathrm{MUCT}(\mathcal{G}, \mathbf{Y})$ and $\mathrm{IB}(\mathcal{G}, \mathbf{Y})$, respectively. Every exogenous variable in $\mathbf{U}$ shall be a binary variable with its domain being $\{0, 1\}$. Let $\oplus$ denote the exclusive-or function and $\bigvee$ the logical OR operator.

(Case $\mathbf{T} = \mathbf{Y}$) In this case, $\mathbf{B}$ corresponds to the parents of $\mathbf{Y}$. Therefore, no target variable Y_i is confounded with $\mathrm{an}(Y_i)_{\mathcal{G}}$ via unobserved confounders. Define an SCM such that

- Each endogenous variable $V \in \mathbf{V}$ is influenced by an exogenous variable $U_V \in \mathbf{V}$;
- $f_{Y_i} = \bigvee \mathbf{u}^{Y_i} \oplus \bigvee \mathbf{pa}_{Y_i}$ with $P(\mathbf{U}^{Y_i} = 0) \approx 1$, for all $i = 1, \ldots, m$;
- $f_X = (\bigoplus \mathbf{u}^X) \oplus (\bigoplus \mathbf{pa}_X)$ for $X \in \mathbf{X}$ and $P(U = 0) = 0.5$ for every $U \in \mathbf{U} \backslash (\bigcup_{i=1}^{m} \mathbf{U}^{Y_i})$.

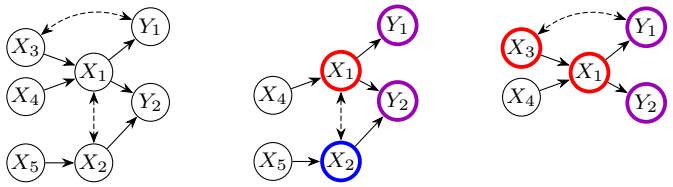
By the third rule of do-calculus and by taking conditional expectations, it holds

$$\begin{aligned}
\mathbb{E}[Y_i | \mathrm{do}(\mathbf{B} = 0)] &= \mathbb{E}[Y_i | \mathrm{do}(\mathrm{pa}(Y_i)_{\mathcal{G}} = 0)] && (4.14)\\
&= \mathbb{E}[Y_i | \mathrm{do}(\mathrm{pa}(Y_i)_{\mathcal{G}} = 0), \mathbf{U}^{Y_i} \neq 0] P(\mathbf{U}^{Y_i} \neq 0) && (4.15)\\
&\quad + \mathbb{E}[Y_i | \mathrm{do}(\mathrm{pa}(Y_i)_{\mathcal{G}} = 0), \mathbf{U}^{Y_i} = 0] P(\mathbf{U}^{Y_i} = 0)\\
&\approx 0 && (4.16)
\end{aligned}$$

for every $1 \leqslant i \leqslant m$. Meanwhile, all other interventions yield expectations greater than or equal to 0.5 in at least one component. Therefore, $\mathbf{B}$ is a possibly-optimal minimal intervention set.

(Case $\mathbf{T} \subset \mathbf{Y}$) In this case, at least one target variable Y_i has an unobserved confounder with its ancestors. As a first step, it will be shown that there exists an SCM, conforming to $\mathcal{H} = \mathcal{G}[\mathbf{T} \cup \mathbf{B}]$, where the intervention $\mathrm{do}(\mathbf{B} = 0)$ is the single best strategy. To achieve this, we first define individual SCMs for each unobserved confounder in $\mathcal{H}[\mathbf{T}]$, and merge them into a single SCM where $\mathrm{do}(\mathbf{B} = 0)$ is indeed the best strategy. Let $\mathbf{U}' = \{U_j\}_{j=1}^{k}$ be the set of unobserved confounders in $\mathcal{H}[\mathbf{T}]$.

Given $U_j \in \mathbf{U}'$, let $B^{(j)}$ and $R^{(j)}$ denote its two children. We define an SCM $\mathcal{M}_j$, where the graph structure is given by



(a) $\mathcal{G}$ (b) $\mathcal{H}_1$ (c) $\mathcal{H}_2$

Figure 4.4: Original causal graph $\mathcal{G}$ and colour-coded graphs for each unobserved confounder.

		$\mathcal{M}_1$				$\mathcal{M}_2$				$\mathcal{M}$			
U_1	U_2	$X_2^{(1)}$	$X_1^{(1)}$	$Y_1^{(1)}$	$Y_2^{(1)}$	$X_3^{(2)}$	$X_1^{(2)}$	$Y_1^{(2)}$	$Y_2^{(2)}$	Y_1'	Y_2'	Y_1	Y_2
0	0	0	1	2	2	1	1	2	2	1010	1010	0	0
0	1					0	0	1	1	0110	0110	0	0
1	0	1	0	1	1	1	1	2	2	1001	1001	0	0
1	1					0	0	1	1	0101	0101	0	0

Table 4.1: Values for $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}$ given $X_4 = X_5 = 0$. The target variables are shown as bit sequences, Y_1' and Y_2', as well as binary values, Y_1 and Y_2.

$$\mathcal{H}_j = \mathcal{H}\left[\mathrm{De}\left(\left\{B^{(j)}, R^{(j)}\right\}\right)_{\mathcal{H}} \cup \left(\mathbf{B} \cap \mathrm{pa}\left(\mathrm{De}\left(\left\{B^{(j)}, R^{(j)}\right\}\right)_{\mathcal{H}}\right)\right)\right], \tag{4.17}$$

and all bidirected edges, except for U_j, are removed. In order to set the structural equations for variables in $\mathcal{H}_j$, the vertices will be labelled via colour coding: Let vertices in $\mathrm{De}\left(B^{(j)}\right)_{\mathcal{H}} \setminus \mathrm{De}\left(R^{(j)}\right)_{\mathcal{H}}$ be labelled as blue, $\mathrm{De}\left(R^{(j)}\right)_{\mathcal{H}} \setminus \mathrm{De}\left(B^{(j)}\right)_{\mathcal{H}}$ as red, and $\mathrm{De}\left(B^{(j)}\right)_{\mathcal{H}} \cap \mathrm{De}\left(R^{(j)}\right)_{\mathcal{H}}$ as purple. All target variables are coloured as purple as well. Moreover, $B^{(j)}$ and $R^{(j)}$ shall perceive U_j as a parent coloured as blue with value U_j and red with value $1 - U_j$, respectively. The blue-, red- and purple-coloured variables are set to 3 if any of their parents in $\mathbf{B}$ is not 0. Otherwise, their values are determined as follows. For every blue and red vertex, the associated structural equation returns the common value of its parents of the same colour and returns 3 if coloured parents' values are not homogeneous. For every purple vertex, its corresponding equation returns 2 if every blue, red and purple parent is 0,1, and 2, respectively, and returns 1 if 1,0,1, respectively.

Next, the SCMs $\mathcal{M}_1, \ldots, \mathcal{M}_k$ will be merged into one single SCM, that conforms to $\mathcal{H}$, and for which $\mathrm{do}(\mathbf{B} = 0)$ is the single best intervention. Note that in $\mathcal{M}_j$ all variables can be represented with just two bits. To construct a unified SCM, variables in $\mathbf{T}$ are represented with $2k$ bits, where $\mathcal{M}_j$ takes the $2j - 1^{\text{th}}$ and $2j^{\text{th}}$ bits. Every target variable Y_i is represented as a sequence of bits and binarised as follows. Y_i is set to 0 if its $2j - 1^{\text{th}}$ and $2j^{\text{th}}$ bits are 00, 01 or 10 for every $1 \leqslant j \leqslant k$, and 1 otherwise. Let $P(U_j = 1) = 0.5$ for $U_j \in \mathbf{U}'$. Therefore, it holds $Y_i = 0$ if $\mathrm{do}(\mathbf{B} = 0)$ and $Y_i = 1$ if $\mathrm{do}(\mathbf{B} \neq 0)$. If any variable in $\mathbf{T}$ is intervened, then at least one SCM $\mathcal{M}_j$ will be disrupted, resulting in an expectation larger than or equal to 0.5 for at least one target variable. In the multi-target setting, it may happen that some target

variables do not occur in any of the $\mathcal{M}_j$'s. This happens if a target Y_i has no parents in $\mathbf{T}$, but only in $\mathbf{B}$. For all such Y_i's, we set $f_{Y_i} = \mathbf{u}^{Y_i} \oplus \bigvee \mathbf{pa}_{Y_i}$ with $P(\mathbf{U}^{Y_i} = 0) \approx 1$. As such, the newly constructed SCM enforces $\mathbb{E}[Y_i|\text{do}(\mathbf{B} = 0)] \approx 0$. Meanwhile, all other interventions yield expectations greater than or equal to 0.5

As a last step, the previously defined SCM for $\mathcal{H} = \mathcal{G}[\mathbf{T} \cup \mathbf{B}]$, will be extended to an SCM for $\mathcal{G}$. However, we can ignore joint probability distributions for any exogenous variables only affecting endogenous variables outside of $\mathcal{H}$. Setting structural equations for endogenous variables outside of $\mathcal{H}$ is redundant as well. For $V \in \text{An}(\mathbf{Y})_{\mathcal{G}} \backslash \mathbf{T}$, we define the structural equations as $f_V = (\bigoplus \mathbf{u}^V) \oplus (\bigoplus \mathbf{pa}_V)$. For $U \in \mathbf{U} \backslash \mathbf{U}'$, we set $P(U = 0) = 0.5$ if U's child(ren) is disjoint to $\mathbf{T}$, and $P(U = 0) \approx 1$ otherwise. Note that $\text{do}(\mathbf{B} = 0)$ is still the single optimal intervention. Therefore, $\mathbf{B}$ is a possibly-optimal minimal intervention set. □

In order to illustrate the construction of an SCM where $\text{do}(\text{IB}(\mathcal{G}, \mathbf{Y}) = 0)$ is the single best strategy, consider Figure 4.4, showing an exemplary graph and its colour-coded subgraphs, $\mathcal{H}_1$ and $\mathcal{H}_2$, for each unobserved confounder. Table 4.1 presents the associates values for $\mathcal{M}_1$ and $\mathcal{M}_2$, as well as values for the target variables in the final SCM $\mathcal{M}$. The next proposition is also a multi-target variant of a result from Lee and Bareinboim, 2018.

Proposition 4.5. *Given* $\langle \mathcal{G}, \mathbf{X}, \mathbf{Y} \rangle$, $\text{IB}(\mathcal{G}_{\overline{\mathbf{X}_s}}, \mathbf{Y})$ *is a possibly-optimal minimal intervention set for any* $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$.

Proof. Let $\mathbf{X}_s$ be an intervention set. Let us denote $\mathbf{T} = \text{MUCT}(\mathcal{G}_{\overline{\mathbf{X}_s}}, \mathbf{Y})$, $\mathbf{B} = \text{IB}(\mathcal{G}_{\overline{\mathbf{X}_s}}, \mathbf{Y})$ and $\mathbf{T}_0 = \text{MUCT}(\mathcal{G}, \mathbf{Y})$. Using the strategy from Proposition 4.5, we construct an SCM for $\mathcal{G}[\mathbf{T} \cup \mathbf{B}]$ while ignoring unobserved confounders between $\mathbf{T}$ and $\mathbf{T}_0 \backslash \mathbf{T}$. Let $\mathbf{U}'$ be the set of such unobserved confounders. Now, the SCM needs to be modified to ensure that $\text{do}(\mathbf{B} = 0)$ is the single best intervention. Every $U \in \mathbf{U}'$ shall flip (i.e., $0 \leftrightarrow 1$) the value of its endogenous child in $\mathbf{T}$ whenever $U = 1$. Let $P(U = 0) \approx 1$, so that it holds $\mathbb{E}[Y_i|\text{do}(\mathbf{B} = 0)] \approx 0$. Intervening on $\mathbf{B} \neq 0$ or on any variable in $\mathbf{T}$ results in expectations around 0.5 or above. □

Notably, Proposition 4.5 extends Proposition 4.4 when $\mathbf{X}_s \neq \varnothing$. See Figure 4.2b, for an example, where $\text{IB}(\mathcal{G}_{\overline{X}_1}, \mathbf{Y}) = \{X_1, X_2\}$ is possibly-optimal. Intervening on X_1 cuts the influence of $\{X_3, X_4\}$ on the targets, and removes the unobserved confounder between X_1 and X_2, while still allowing it to affect X_2. Moreover, in Figure 4.2c it holds $\text{IB}(\mathcal{G}_{\overline{X}_3}, \mathbf{Y}) = \{X_1, X_2\}$ because intervening on X_3 removes the unobserved confounder between X_3 and Y_1. Note that, by iterating over all intervention sets $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$, we can discover possibly-optimal minimal intervention sets in a given graph. The following theorem is an extension of the main result by Lee and Bareinboim, 2018 to the scenario where multiple target variables are present. It shows that the aforementioned strategy suffices to find not some, but all, possibly-optimal minimal intervention sets for $\langle \mathcal{G}, \mathbf{X}, \mathbf{Y} \rangle$.

Theorem 4.1. *Given* $\langle \mathcal{G}, \mathbf{X}, \mathbf{Y} \rangle$, *an intervention set* $\mathbf{X}_s$ *is possibly-optimal if and only if it holds* $\text{IB}(\mathcal{G}_{\overline{\mathbf{X}_s}}, \mathbf{Y}) = \mathbf{X}_s$.

Proof. (If) This is a special case of Proposition 4.5.

(Only if) Let $\mathbf{X}_s$ be a minimal intervention set and $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ an intervention value. Denote $\mathbf{T} = \text{MUCT}(\mathcal{G}_{\overline{\mathbf{X}}_s}, \mathbf{Y})$, $\mathbf{B} = \text{IB}(\mathcal{G}_{\overline{\mathbf{X}}_s}, \mathbf{Y})$, $\mathbf{T}_0 = \text{MUCT}(\mathcal{G}, \mathbf{Y})$ and $\mathbf{B}_0 = \text{IB}(\mathcal{G}, \mathbf{Y})$. From Proposition 4.3, we know that no possibly-optimal minimal intervention set intersects with $\text{An}(\mathbf{B}_0)_{\mathcal{G}} \backslash \mathbf{B}_0$ and thus, it is possible to conclude $\mathbf{X}_s \subseteq \mathbf{T}_0 \cup \mathbf{B}_0 \backslash \mathbf{Y}$. The strategy is now to prove that $\mathbf{X}_s$ cannot be possibly-optimal whenever $\mathbf{X}_s \neq \mathbf{B}$. Assume that $\mathbf{X}_s \neq \mathbf{B}$ and let $\mathbf{B}' = \mathbf{B} \backslash (\mathbf{X}_s \cap \mathbf{B})$. Note that it holds $\mathbf{X}_s \subseteq \text{An}(\mathbf{B})_{\mathcal{G}}$ since otherwise it would follow $\mathbf{X}_s \cap \mathbf{T} \neq \varnothing$, which contradicts that $\mathbf{X}_s$ is neither a descendant of some variable nor confounded in $\mathcal{G}_{\overline{\mathbf{X}}_s}$. Then, for all $i = 1, \ldots, m$ it holds

$$\mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x_s})] = \int_{\mathcal{D}(\mathbf{B}')} \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x_s}), \mathbf{B}' = \mathbf{b}'] P(\mathbf{B}' = \mathbf{b}'|\text{do}(\mathbf{X}_s = \mathbf{x_s})) d\mathbf{b}' \tag{4.18}$$

$$= \int_{\mathcal{D}(\mathbf{B}')} \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x_s}, \mathbf{B}' = \mathbf{b}')] P(\mathbf{B}' = \mathbf{b}'|\text{do}(\mathbf{X}_s = \mathbf{x_s})) d\mathbf{b}' \tag{4.19}$$

$$= \int_{\mathcal{D}(\mathbf{B}')} \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s \cap \mathbf{B} = \mathbf{x_s}[\mathbf{B}], \mathbf{B}' = \mathbf{b}')] P(\mathbf{B}' = \mathbf{b}'|\text{do}(\mathbf{X}_s = \mathbf{x_s})) d\mathbf{b}', \tag{4.20}$$

where the second and third equalities result from the second and third rules of do-calculus, respectively. Similarly to the proofs of Proposition 4.2 and Proposition 4.3, we can show via contradiction that $\mathbf{X}_s$ cannot be a possibly-optimal minimal intervention set. □

Theorem 4.1 provides a necessary and sufficient condition for a set of variables to be a possibly-optimal minimal intervention set, given $\langle \mathcal{G}, \mathbf{X}, \mathbf{Y} \rangle$. For instance, in Figure 4.2d, it holds $\text{IB}(\mathcal{G}_{\overline{\{X_2, X_4\}}}, \mathbf{Y}) = \{X_2, X_4\}$, and therefore $\{X_2, X_4\}$ is possibly-optimal. Moreover, $\text{pa}(\mathbf{Y})_{\mathcal{G}}$ is always a possibly-optimal minimal intervention set since $\text{MUCT}(\mathcal{G}_{\overline{\text{pa}(\mathbf{Y})_{\mathcal{G}}}}, \mathbf{Y}) = \mathbf{Y}$ and therefore $\text{IB}(\mathcal{G}_{\overline{\text{pa}(\mathbf{Y})_{\mathcal{G}}}}, \mathbf{Y}) = \text{pa}(\mathbf{Y})_{\mathcal{G}}$.

In this chapter, we studied the problem of which variables to intervene upon in order to establish dominant trade-offs between the target variables in a given causal graph. Building on existing literature on the topic (Lee & Bareinboim, 2018), we extended the analysis to the multi-objective setting. Collectively, these results culminate in the characterisation of the possibly-optimal minimal intervention sets, addressing the problem statement outlined in section 4.1.

5 Multi-objective causal Bayesian optimisation (MO-CBO)

Drawing from methods in multi-objective optimisation, we extend causal Bayesian optimisation (CBO) for causal models featuring multiple outcome variables, thereby introducing *multi-objective causal Bayesian optimisation* (MO-CBO). A MO-CBO problem involves sequentially performing interventions in a causal system in order to optimise multiple outcome variables simultaneously, subject to a budgetary constraint as interventions are assumed to be costly. We propose a novel algorithm, called CAUSAL PARETOSELECT, that solves MO-CBO problems given a collection of intervention sets to be considered in the optimisation procedure. To this end, our methodology introduces the notion of so called *causal Pareto fronts*, where the corresponding Pareto set consists of solutions in terms of intervention set-value pairs. To the best of our knowledge, this is the first research effort towards MO-CBO. In summary, the contributions of this chapter are as follows:

1. In the next section, we formalise a new class of optimisation problems called *multi-objective causal Bayesian optimisation* (MO-CBO), where the aim is to sequentially intervene on a causal system in order identify non-dominated trade-offs between multiple output variables (section 5.1).

2. In order to solve MO-CBO problems, we introduce the notion of *causal Pareto fronts*. We will show that the problem of discovering the causal Pareto front can be decomposed into several classical multi-objective optimisation problems (section 5.2).

3. We propose an algorithm towards solving MO-CBO problems, called CAUSAL PARETOSELECT. It builds on ideas from multi-objective Bayesian optimisation (MOBO) and intents to construct the causal Pareto front (section 5.4). We integrate a Gaussian process surrogate to model the interventional effects. Moreover, we define a custom acquisition function, called *relative hypervolume improvement*, which drives the exploration of different intervention sets. The code is available at: `https://github.com/ShriyaBhatija/Causal-ParetoSelect`.

5.1 MO-CBO formalism

This section formally introduces the problem formulation of multi-objective causal Bayesian optimisation along with challenges that may arise.

5.1.1 Problem setting

Let $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ be a structural causal model (SCM) and $\mathcal{G}$ its associated graph that encodes the underlying causal mechanisms. Recall that $\mathbf{U}$ is a set of independent exogenous variables distributed according to the probability distribution $P(\mathbf{U})$, $\mathbf{V}$ is a set of endogenous variables, and $\mathbf{F} = \{f_V\}_{V \in \mathbf{V}}$ is a set of functions such that $V = f_V(\text{pa}(V)_{\mathcal{G}}, \mathbf{U}^V)$. Within $\mathbf{V}$ there are two different types of variables to be distinguished: treatment variables $\mathbf{X}$, that can be set to specific values, and output variables $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$, that represent the outcome of interest and are non-manipulative. A set $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ is called an *intervention set* and $\mathbf{x}_s \in \mathcal{D}(\mathbf{X}_s)$ an *intervention value*. The tuple $(\mathbf{X}_s, \mathbf{x}_s)$ is referred to as *intervention set-value* pair.

5.1.2 Formulation of MO-CBO

In MO-CBO problems, the aim is to identify intervention set-value pairs $(\mathbf{X}_s, \mathbf{x}_s)$ that can minimise (or, for the dual problem, maximise) all target variables $Y_1, \ldots, Y_m$ simultaneously. The outcomes of an intervention $\text{do}(\mathbf{X}_s = \mathbf{x}_s)$ are captured as the expected values of each target:

$$\mathbb{E}_{P(Y_i|\text{do}(\mathbf{X}_s=\mathbf{x}_s))}[Y_i] := \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)], \tag{5.1}$$

where $P(Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s))$ denotes the interventional distribution of Y_i, for all $i = 1, \ldots, m$. We write $\mu_i(\mathbf{X}_s, \mathbf{x}_s) = \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$, and $\boldsymbol{\mu}(\mathbf{X}_s, \mathbf{x}_s) = (\mu_1(\mathbf{X}_s, \mathbf{x}_s), \ldots, \mu_m(\mathbf{X}_s, \mathbf{x}_s))$ for the vector notation. The objective functions of a MO-CBO problem are thus given by $\mu_i : \mathbb{X} \to \mathbb{R}$, $1 \leqslant i \leqslant m$, where $\mathbb{X}$ is the space of all intervention set-value pairs. The averaged outcomes are computed according to the interventional distribution, inherently aiming to isolate cause-effect relations. By taking expectations, the μ_i's account for the uncertainty proposed by the exogenous variables in $\mathbf{U}$. Further, note that the objectives can be conflicting due to possibly opposing causal relationships among the variables, which motivates looking at multi-objective optimisation settings, where the aim is to find Pareto-optimal solutions that yield the best trade-offs between objective functions. This motivates our adaptation of traditional Pareto optimality (Miettinen, 1999) to intervention set-value pairs, as given in the definitions below.

Definition 5.1 (Pareto-optimal intervention set-value pairs)**.** Given $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, an intervention set-value pair $(\mathbf{X}_s, \mathbf{x}_s)$ with $\mathbf{X}_s \in \mathcal{S}$ is called *Pareto-optimal for* $\mathcal{S}$, or else *non-dominated for* $\mathcal{S}$, if there is no other intervention set-value pair $(\mathbf{X}'_s, \mathbf{x}'_s)$ with $\mathbf{X}'_s \in \mathcal{S}$ such that $\mu_i(\mathbf{X}_s, \mathbf{x}_s) \geqslant \mu_i(\mathbf{X}'_s, \mathbf{x}'_s)$ for all $1 \leqslant i \leqslant m$, and $\mu_i(\mathbf{X}_s, \mathbf{x}_s) > \mu_i(\mathbf{X}'_s, \mathbf{x}'_s)$ for at least one $1 \leqslant i \leqslant m$.

Definition 5.2 (Causal Pareto set and causal Pareto front)**.** Given $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, the space of all Pareto-optimal intervention set-value pairs for $\mathcal{S}$ is called *causal Pareto set for* $\mathcal{S}$ and shall be denoted $\mathcal{P}^{\mathcal{C}}_s(\mathcal{S})$. The corresponding *causal Pareto front for* $\mathcal{S}$, denoted $\mathcal{P}^{\mathcal{C}}_f(\mathcal{S})$, is the m-dimensional image of the causal Pareto set under the objectives μ_i, $1 \leqslant i \leqslant m$.

Problem statement. A MO-CBO problem involves identifying solutions in $\mathcal{P}^{\mathcal{C}}_f(\mathcal{P}(\mathbf{X}))$, and we refer to $\mathcal{P}^{\mathcal{C}}_f(\mathcal{P}(\mathbf{X}))$ simply as *the causal Pareto front*.

5.1.3 Challenges

The above defined optimisation problem proposes several challenges, of which the latter two were directly inherited from CBO and already mentioned by Aglietti et al., 2020:

Complexity of the design space. Firstly, traditional Pareto optimality is defined in the context of multi-objective optimisation problems where the design space is typically a subset of $\mathbb{R}^d$, $d \geqslant 2$. In contrast, the optimisation problem here aims to identify optimal trade-offs with respect to a set of intervention set-value pairs. The challenge lies in navigating optimisation in this non-trivial design space.

Exponantial growth of the search space. Secondly, finding the best trade-offs would, in principle, require exploring the space of all intervention sets $\mathcal{P}(\mathbf{X})$ which grows exponentially in cardinality with respect to the number of treatment variables. This makes the search for Pareto-optimal intervention set-value pairs computationally expensive and potentially infeasible for large-scale systems.

Cost of interventions. Thirdly, the proposed framework necessitates evaluating the objective functions, which inherently involves implementing and observing the effects of interventions on the system. However, in practical scenarios, performing interventions is often expensive and thus, it is crucial to ensure sample efficiency and account for the cost $co(\mathbf{X}_s, \mathbf{x}_s)$ of an intervention $\text{do}(\mathbf{X}_s = \mathbf{x}_s)$.

All of these challenges will be addressed with in the upcoming sections, and our algorithm CAUSAL PARETOSELECT is specifically designed to overcome them.

5.2 Causal Pareto front and local MO-CBO problems

This section addresses the first of the aforementioned challenges, namely searching for Pareto-optimal solutions within the non-trivial design space of MO-CBO problems, consisting of intervention set-value pairs. We first introduce so called *local MO-CBO problems*:

Definition 5.3 (Local MO-CBO problem)**.** Let $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ be an intervention set. Then, the multi-objective optimisation problem defined by the functions $\mu_i(\mathbf{X}_s, \cdot) : \mathcal{D}(\mathbf{X}_s) \to \mathbb{R}$, $\mathbf{x}_s \mapsto \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$, $1 \leqslant i \leqslant m$, is referred to as *local mo-cbo problem w.r.t.* $\mathbf{X}_s$.

For a local MO-CBO problem w.r.t. an intervention set $\mathbf{X}_s$, its Pareto set is denoted as $\mathcal{P}_s(\mathbf{X}_s)$ and the associated Pareto front is denoted as $\mathcal{P}_f(\mathbf{X}_s)$. Note that each local MO-CBO problem corresponds to a standard multi-objective optimisation task and established algorithms can be employed to compute the solutions for these problems.

Let $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$ be a collection of intervention sets. Let $\mathcal{P}_s^{\mathcal{C}}(\mathcal{S})$ denote the causal Pareto set for $\mathcal{S}$ and $\mathcal{P}_f^{\mathcal{C}}(\mathcal{S})$ its associated causal Pareto front for $\mathcal{S}$. The following proposition shows that

the causal Pareto front is contained within the solutions of the local MO-CBO problems.

Proposition 5.1 (MO-CBO decomposition)**.** *Let $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$ be a collection of intervention sets and $\mathcal{P}_f(\mathbf{X}_1), \dots, \mathcal{P}_f(\mathbf{X}_{|\mathcal{S}|})$ the Pareto fronts of the associated local MO-CBO problems. Then, it holds*

$$\mathcal{P}_f^{\mathcal{C}}(\mathcal{S}) \subseteq \bigcup_{s=1}^{|\mathcal{S}|} \mathcal{P}_f(\mathbf{X}_s). \tag{5.2}$$

Proof. Assume for contradiction that $\mathcal{P}_f^{\mathcal{C}}(\mathcal{S}) \nsubseteq \bigcup_{s=1}^{|\mathcal{S}|} \mathcal{P}_f(\mathbf{X}_s)$. Then, there exists some $\mathbf{z} \in \mathbb{R}^m$ such that $\mathbf{z} \in \mathcal{P}_f^{\mathcal{C}}(\mathcal{S})$ and $\mathbf{z} \notin \mathcal{P}_f(\mathbf{X}_s)$ for all $s = 1, \dots, |\mathcal{S}|$. For some intervention set $\mathbf{X}'_s \in \{\mathbf{X}_1, \dots, \mathbf{X}_{|\mathcal{S}|}\}$ and intervention value $\mathbf{x}'_s \in \mathcal{D}(\mathbf{X}'_s)$, it holds $\mathbf{z} = (\mu_1(\mathbf{X}'_s, \mathbf{x}'_s), \dots, \mu_m(\mathbf{X}'_s, \mathbf{x}'_s))$. Let $\mathcal{P}_s(\mathbf{X}_1), \dots, \mathcal{P}_s(\mathbf{X}_{|\mathcal{S}|})$ be the Pareto sets of the associated local MO-CBO problems w.r.t. $\mathbf{X}_1, \dots, \mathbf{X}_{|\mathcal{S}|}$. Since $\mathbf{z} \notin \mathcal{P}_f(\mathbf{X}'_s)$, it follows $\mathbf{x}'_s \notin \mathcal{P}_s(\mathbf{X}'_s)$, i.e. $\mathbf{x}'_s$ is not Pareto-optimal in the local MO-CBO problem w.r.t. $\mathbf{X}'_s$. Thus, there exists another intervention value $\mathbf{x}'' \in \mathcal{D}(\mathbf{X}'_s)$ such that $\mu_i(\mathbf{X}'_s, \mathbf{x}'_s) \geqslant \mu_i(\mathbf{X}'_s, \mathbf{x}''_s)$ for all i and $\mu_i(\mathbf{X}'_s, \mathbf{x}'_s) > \mu_i(\mathbf{X}'_s, \mathbf{x}''_s)$ for at least one i, $i = 1, \dots, m$. In other words, the intervention set-value pair $(\mathbf{X}'_s, \mathbf{x}'_s)$ is not Pareto-optimal since it is dominated by $(\mathbf{X}'_s, \mathbf{x}''_s)$. Therefore, $\mathbf{z} \notin \mathcal{P}_f^{\mathcal{C}}(\mathbf{X}_1, \dots, \mathbf{X}_{|\mathcal{S}|})$ which is a contradiction. □

In other words, for a given collection of intervention sets $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, the causal Pareto front $\mathcal{P}_f^{\mathcal{C}}(\mathcal{S})$ lies within the Pareto fronts of the local MO-CBO problems w.r.t. $\mathbf{X}_1, \dots, \mathbf{X}_{|\mathcal{S}|}$. Therefore, the problem of discovering the causal Pareto front reduces to identifying the non-dominated solutions within the Pareto fronts of the local MO-CBO problems in $\mathcal{P}(\mathbf{X})$.

5.3 Reducing the search space

This section addresses one of the challenges discussed in section 5.1, specifically the complexity and cost in exploring $2^{|\mathbf{X}|}$ intervention sets in order to identify solutions in the causal Pareto front for $\mathcal{P}(\mathbf{X})$, and solve MO-CBO problems. Albeit, such a brute-force exploration is indeed a valid strategy, the question arises whether there are intervention sets that are redundant to consider, given the topology of $\mathcal{G}$. Then, it would only require exploring a subset $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, referred to as *exploration set*. This idea is formalised in chapter 4 where we exploit the rules of do-calculus to identify invariances among intervention sets as well as discover intervention sets that are possibly-optimal in regards to yielding optimal trade-offs between the targets. The precise methodology in chapter 4 builds on the work by Lee and Bareinboim, 2018 and extends it to the multi-objective setting. We argue that setting $\mathcal{S} = \mathbb{P}_{\mathcal{G},\mathbf{Y}}$ as the the exploration sets suffices for solving MO-CBO problems, albeit the next sections are agnostic to the choice of $\mathcal{S}$.

5.4 Solving the local MO-CBO problems

We have discussed the theoretical notions relevant for MO-CBO in chapter 4 and section 5.2, building a coherent mathematical framework around its formulation. This section synthesises

these ideas and presents a novel methodology for solving MO-CBO problems. Recall that addressing MO-CBO problems requires discovering solutions in the causal Pareto front which, as demonstrated in section 5.2, consists of non-dominated points within the union of Pareto fronts from local MO-CBO problems. Consequently, the task for solving MO-CBO problems is to solve these local problems while keeping the number of function evaluations low. We propose an algorithm that can explore all (possibly-optimal) minimal intervention sets, inherently leveraging the known causal graph structure. This approach not only enhances computational efficiency but also directs exploration towards those local problems for which the Pareto fronts offer the highest potential for improvement. Notably, CAUSAL PARETOSELECT is an extension of a multi-objective Bayesian (MOBO) algorithm, and thus requires an appropriate surrogate model, acquisition function, and Pareto front approximation method. To this end, we consider well established MOBO algorithms that have demonstrated effective strategies. Specifically, DGEMO (Konakovic Lukovic et al., 2020) employs a novel batch selection strategy that optimises for both hypervolume improvement and sample diversity. The authors validate their approach with experiments on both synthetic test functions and real-world benchmark problems. It is shown that DGEMO approaches/outperforms many established MOBO algorithms in terms of prediction error and especially sample efficiency. In the following sections, we discuss our choice of surrogate model, acquisition function, and Pareto front approximation method, which are partially based on Konakovic Lukovic et al., 2020.

Let us fix an exploration set $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, where reasonable choices include $\mathcal{S} \in \{\mathbb{M}_{\mathcal{G},\mathbf{Y}}, \mathbb{P}_{\mathcal{G},\mathbf{Y}}\}$ as discussed in chapter 4.

5.4.1 Surrogate model

Following DGEMO (Konakovic Lukovic et al., 2020), MO-CBO will implement a Gaussian process surrogate model that estimates causal effects from interventions. For all $\mathbf{X}_s \in \mathcal{S}$, we consider their respective local MO-CBO problem and associated objective functions given by $\mu_i(\mathbf{X}_s, \cdot) : \mathcal{D}(\mathbf{X}_s) \to \mathbb{R}$, $\mathbf{x}_s \mapsto \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$, for $i = 1, \dots, m$. Here, each objective will be modelled independently, and we denote the prior distribution for a given input $\mathbf{x}_s$ as $\mu_i(\mathbf{X}_s, \mathbf{x}_s) \sim \mathcal{GP}(m_i(\mathbf{x}_s), k_i(\mathbf{x}_s, \mathbf{x}'_s))$, where m_i and k_i are the mean and covariance functions, respectively. A Gaussian process prior can incorporate prior knowledge about the objective functions (if available) through appropriate choices of the mean and kernel functions. We assume no prior knowledge and therefore set $m_i(\mathbf{x}_s) = 0$. Moreover, so called *Matérn kernels functions* (Williams & Rasmussen, 2006) will be used as the covariance function since they can model a large range of function properties. The prior function is refined by training the posterior using a so called *interventional dataset* $\mathcal{D}^I = \{(\mathbf{x}_s^j, \mu(\mathbf{X}_s, \mathbf{x}_s^j)\}_{j=1}^N$. The Gaussian process parameters θ (e.g., $\theta = \ell$ for the Matérn kernel) are optimised by maximising the log marginal likelihood

$$\log\left(p\left(\left(\mu(\mathbf{X}_s, \mathbf{x}_s^1), \dots, \mu(\mathbf{X}_s, \mathbf{x}_s^N)\right)^T \mid \left(\mathbf{x}_s^1, \dots, \mathbf{x}_s^N\right)^T, \theta\right)\right) \tag{5.3}$$

For more details on the surrogate model as well as Matérn kernel, we refer to Konakovic Lukovic et al., 2020 or (Williams & Rasmussen, 2006).

5.4.2 MOBO solver

In this section, we discuss DGEMO's Pareto front approximation algorithm as well as batch selection strategy with respect to solving the local MO-CBO problems in $\mathcal{S}$. For an intervention set $\mathbf{X}_s \in \mathcal{S}$, consider its respective local MO-CBO problem, characterised by the objective functions $\mu_i(\mathbf{X}_s, \cdot) : \mathcal{D}(\mathbf{X}_s) \to \mathbb{R}, \mathbf{x}_s \mapsto \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$ for $i = 1, \ldots, m$. The corresponding Pareto front $\mathcal{P}_f(\mathbf{X}_s)$ will be computed over the mean of the Gaussian process posteriors of the $\mu_i(\mathbf{X}_s, \cdot)$, $1 \leqslant i \leqslant m$. More specifically, DGEMO relies on the approximation approach introduced by Schulz et al., 2018, which derives a first-order approximation of the Pareto front by solving a dual problem based on the *KKT conditions* (Hillermeier, 2001). Once a single Pareto-optimal point is found, the method efficiently uncovers larger Pareto-optimal regions in its neighbourhood. Note that this approach relies on the assumption that the objective functions are continuously differentiable and non-negative. More details on this approach are provided by Schulz et al., 2018.

5.4.3 Acquisition function

The batch selection strategy determines how to select a batch of samples that is to be evaluated in the next iteration. Konakovic Lukovic et al., 2020 propose with DGEMO a novel method that considers sample diversity in the input space. For this intent, they define the so called *diversity regions*: Given a Pareto front approximation, the aim is to group the optimal points based on their performance as well as properties in the design space. From the given Pareto front representation, it is possible to trace along neighbouring points and their correlation in both design and performance space. To address the grouping problem, DGEMO relies on a data structure called *performance buffer* implemented by Schulz et al., 2018. The performance buffer filters the best performing points from the Pareto front approximation by storing them into an $(m-1)$-dimensional array discretised by hyperspherical coordinates. Assuming that all objectives are positive, any point lying on the Pareto front will intersect a positive ray traced from the origin. The best performing points are selected as the ones closest to the origin. Moreover, the buffer stores the locations of the selected points in both performance and design space, as well as the corresponding linear subspace of the design space extracted with the first-order approximation step. Note that the buffer is compatible with a graph-cut method, which can extract a sparse subset of optimal points grouped into K linear subspaces. Such subspaces capture the performance quality as well as neighbourhoods in the design space. The aforementioned diversity regions are precisely these linear subspaces. Next, it is discussed how these diversity regions can facilitate the diversity among samples.

For $\mathbf{X}_s \in \mathcal{S}$, consider its respective local MO-CBO problem and associated objective functions $\mu_i(\mathbf{X}_s, \cdot) : \mathcal{D}(\mathbf{X}_s) \to \mathbb{R}_{\geqslant 0}, \mathbf{x}_s \mapsto \mathbb{E}[Y_i|\text{do}(\mathbf{X}_s = \mathbf{x}_s)]$, for $i = 1, \ldots, m$. Let $\tilde{\mathcal{P}}_f(\mathbf{X}_s)$ be the current Pareto front approximation and let $\mathcal{R}_1(\mathbf{X}_s), \ldots, \mathcal{R}_K(\mathbf{X}_s)$ denote the identified diversity regions. Recall from section 2.1 that to measure the quality of an approximated Pareto front $\tilde{\mathcal{P}}_f(\mathbf{X}_s)$, the hypervolume indicator $\mathcal{H}(\tilde{\mathcal{P}}_f(\mathbf{X}_s))$ (Zitzler & Thiele, 1999) is the most commonly used metric in multi-objective optimisation (Riquelme et al., 2015). The higher the hyper-

volume, the better $\tilde{\mathcal{P}}_f(\mathbf{X}_s)$ approximates the true Pareto front. To determine how much the hypervolume would increase if a batch of samples $\mathbf{B}_s = \{\mathbf{x}_s^1, \dots, \mathbf{x}_s^B\}$ is added to the current Pareto front approximation, hypervolume improvement is used:

$$\text{HVI}(\mathbf{B}_s, \tilde{\mathcal{P}}_f(\mathbf{X}_s)) = \mathcal{H}(\tilde{\mathcal{P}}_f(\mathbf{X}_s) \cup \mathbf{B}_s) - \mathcal{H}(\tilde{\mathcal{P}}_f(\mathbf{X}_s)). \tag{5.4}$$

When selecting the next batch of samples to evaluate, DGEMO considers hypervolume improvement as well as sample diversity. Formally, a batch is chosen as follows:

$$\mathbf{B}_s = \underset{\mathbf{B}_s \subseteq \mathcal{D}(\mathbf{X}_s), |\mathbf{B}_s| = B}{\arg\max} \text{HVI}(\boldsymbol{\mu}(\mathbf{X}_s, \mathbf{B}_s), \tilde{\mathcal{P}}_f(\mathbf{X_s})) \text{ s.t. } \max_{1 \leqslant k \leqslant K} \delta_k(\mathbf{B}_s) - \min_{1 \leqslant k \leqslant K} \delta_k(\mathbf{B}_s) \leqslant 1, \tag{5.5}$$

where the functions $\delta_k(\cdot)$ are defined as the number of elements from $\mathbf{B}_s$ that belong to $\mathcal{R}_k(\mathbf{X}_s)$. Instead of solving this with a combinatorial approach, DGEMO was designed to reduce computational load by employing a greedy approach. For the complete selection algorithm, see (Konakovic Lukovic et al., 2020).

Causal ParetoSelect seeks to balance the exploration of Pareto fronts associated with different local MO-CBO problems and thus, evaluating all $\mathbf{B}_s$, $s = 1, \dots, |\mathcal{S}|$, during a single iteration, is an inefficient strategy. Rather, at each iteration, we select the batch with the most promising hypervolume improvement. For this intent, we introduce a novel acquisition function based on of *relative hypervolume improvement*, defined as:

$$\text{RHVI}(\boldsymbol{\mu}(\mathbf{X}_s, \mathbf{B}_s), \tilde{\mathcal{P}}_f(\mathbf{X_s})) = \text{HVI}(\boldsymbol{\mu}(\mathbf{X}_s, \mathbf{B}_s), \tilde{\mathcal{P}}_f(\mathbf{X_s})) \;/\; \mathcal{H}(\tilde{\mathcal{P}}_f(\mathbf{X}_s)). \tag{5.6}$$

As the name suggests, relative hypervolume improvement is a normalised measure of improvement and therefore provides a more accurate assessment when comparing batches across different intervention sets. Given $\mathbf{B}_1, \dots, \mathbf{B}_{|\mathcal{S}|}$, we propose the following batch selection strategy for Causal ParetoSelect:

$$\mathbf{B}_{eval} = \underset{\mathbf{B}_s \in \{\mathbf{B}_1, \dots, \mathbf{B}_{|\mathcal{S}|}\}}{\arg\max} \text{RHVI}(\boldsymbol{\mu}(\mathbf{X}_s, \mathbf{B}_s), \tilde{\mathcal{P}}_f(\mathbf{X_s})). \tag{5.7}$$

This addresses one of challenges discussed in section 5.1, namely that interventions are expensive to perform and it is desirable to minimise these costs. The proposed batch selection is designed to offer a trade-off between advancing the Pareto fronts $\mathcal{P}(\mathbf{X}_1), \dots, \mathcal{P}(\mathbf{X}_{|\mathcal{S}|})$ while simultaneously keeping the number of interventions low.

5.5 MO-CBO algorithm

In this section, we present a novel algorithm for addressing MO-CBO, called Causal ParetoSelect. The complete procedure is outlined in algorithm 3. Given an exploration set $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, we assume the availability of prior interventional data for all $\mathbf{X}_s \in \mathcal{S}$. This dataset is called the *interventional dataset* and is denoted as $\mathcal{D}^I = \{((\mathbf{X}_s, \mathbf{x}_s^k), \boldsymbol{\mu}(\mathbf{X}_s, \mathbf{x}_s^k))\}_{k=1,s=1}^{K,|\mathcal{S}|}$, where K represents the number of intervention values per set.

Algorithm 3: CAUSAL PARETOSELECT (CPS)

Input: Causal graph $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X} \rangle$, exploration set $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, local MO-CBO problems $\mu_i(\mathbf{X}_s, \cdot) : \mathcal{D}(\mathbf{X}_s) \to \mathbb{R}$ for $\mathbf{X}_s \in \mathcal{S}$, interventional data $\mathcal{D}^I$, batch size B, number of iterations N

Output: Causal Pareto set $\mathcal{P}_s^{\mathcal{C}}(\mathcal{S})$ and causal Pareto front $\mathcal{P}_f^{\mathcal{C}}(\mathcal{S})$

Initialise the dataset $\mathcal{D}_0^I = \mathcal{D}^I$

for $s = 1, \ldots, |\mathcal{S}|$ **do**

  Fit surrogate models $\tilde{\mu}_i(\mathbf{X_s}, \cdot)$ using $\mathcal{D}_0^I$, for all $i = 1, \ldots, m$

  Approximate Pareto set $\tilde{\mathcal{P}}_s(\mathbf{X}_s)$ and Pareto front $\tilde{\mathcal{P}}_f(\mathbf{X}_s)$ using $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$

for $n = 1, \ldots, N$ **do**

  for $s = 1, \ldots, |\mathcal{S}|$ **do**

    Select batch of samples $\mathbf{B}_s = \{\mathbf{x}_s^b\}_{b=1}^B$ via equation (5.5)

  Select batch of samples $\mathbf{B}_{\hat{s}}$ from $\{\mathbf{B}_1, \ldots, \mathbf{B}_{|\mathcal{S}|}\}$ via equation (5.7)

  Intervene on $\mathbf{X}_{\hat{s}}$ with $\mathbf{B}_{\hat{s}}$

  Augment $\mathcal{D}_n^I = \mathcal{D}_{n-1}^I \cup \{(\mathbf{X}_{\hat{s}}, \mathbf{x}_{\hat{s}}^b), \boldsymbol{\mu}(\mathbf{X}_{\hat{s}}, \mathbf{x}_{\hat{s}}^b))\}_{b=1}^B$

  Update surrogate models $\tilde{\mu}_i(\mathbf{X}_{\hat{s}}, \cdot)$ using $\mathcal{D}_n^I$, for all $i = 1, \ldots, m$

  Approximate Pareto set $\tilde{\mathcal{P}}_s(\mathbf{X}_{\hat{s}})$ and Pareto front $\tilde{\mathcal{P}}_f(\mathbf{X}_{\hat{s}})$ using $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$

Compute $\mathcal{P}_s(\mathbf{X}_1), \ldots, \mathcal{P}_s(\mathbf{X}_{|\mathcal{S}|})$ and $\mathcal{P}_f(\mathbf{X}_1), \ldots, \mathcal{P}_f(\mathbf{X}_{|\mathcal{S}|})$ from $\mathcal{D}_N^I$

Compute causal Pareto set $\mathcal{P}_s^{\mathcal{C}}(\mathcal{S})$ and causal Pareto front $\mathcal{P}_f^{\mathcal{C}}(\mathcal{S})$

CAUSAL PARETOSELECT addresses the MO-CBO problem from section 5.1 by iteratively advancing the Pareto fronts of each local MO-CBO problem in the exploration set $\mathcal{S}$. Initially, for each intervention set, we fit the surrogate model with the initial interventional data $\mathcal{D}_0^I$, and approximate the Pareto sets and fronts of all local MO-CBO problems given the fitted surrogates. Thereafter, at each iteration, we select the most promising intervention set in regards to relative hypervolume improvement for batch evaluation. The dataset is augmented with the newly evaluated batch of samples. For the corresponding intervention set, we again update the surrogate model, as well as Pareto set and front approximations. This procedure is repeated for a specified number of iterations. After completing all iterations, the algorithm computes the Pareto sets and fronts for each local MO-CBO problem using the collected objective function evaluations. The causal Pareto set $\mathcal{P}_s^{\mathcal{C}}(\mathcal{S})$ and causal Pareto front $\mathcal{P}_f^{\mathcal{C}}(\mathcal{S})$ are subsequently identified by determining the non-dominated points across the local problems.

5.6 Runtime complexity

The time complexity of CAUSAL PARETOSELECT can be divided into several stages: surrogate model fitting, Pareto set and Pareto front approximations, batch selection for each intervention set in $\mathcal{S}$, and overall batch selection with relative hypervolume improvement. We discuss the

complexity for each component of CAUSAL PARETOSELECT given the analysis from DGEMO (Konakovic Lukovic et al., 2020).

1. Surrogate model fitting. Fitting the surrogate model for a selected local MO-CBO problem $\mathbf{X}_s \in \mathcal{S}$ has a time complexity of $\mathcal{O}(N_s^3)$, where N_s is the size of the current interventional dataset w.r.t $\mathbf{X}_s$. Note that the selected local MO-CBO problem as well as dataset size change throughout iterations $n = 1, \ldots, N$.

2. Pareto set and Pareto front approximations. The complexity for approximating the Pareto sets and fronts for the local MO-CBO problem w.r.t. $\mathbf{X}_s \in \mathcal{S}$ is $\mathcal{O}(N_s^{\text{gen}} N_s^{\text{pop}} (d_s^3 + N_s n_s))$, where N_s^{gen} and N_s^{pop} represent the number of generations and population size, respectively. For more details on these terms from multi-objective optimisation we refer to (Konakovic Lukovic et al., 2020). Moreover, d_s is the dimension of the input variables, N_s the number of interventional data points, and n_s number of grid samples, all with respect to $\mathbf{X}_s$.

3. Batch selection for local MO-CBO problems. For each local MO-CBO problem w.r.t $\mathbf{X}_s \in \mathcal{S}$, batch selection is performed by computing the hypervolume improvement for each candidate batch. The time complexity of batch selection depends on the number of objectives m and is given as:

- For $m = 2$, the complexity is $\mathcal{O}(BDN_s^{\text{pf}})$,
- For $m = 3$, the complexity is $\mathcal{O}(BDN_s^{\text{pf}} \log N_{\text{pf}})$,
- For $m > 3$, the complexity is $\mathcal{O}(BD(N_s^{\text{pf}})^{\lceil (m-1)/2 \rceil + 1})$,

where B is the batch size, D is the maximum performance buffer capacity (Konakovic Lukovic et al., 2020), and N_s^{pf} is the number of points on the current Pareto front.

4. Batch selection between local MO-CBO problems. After obtaining the candidate batches from each local MO-CBO problem, the batch with the highest relative hypervolume improvement is selected. This step involves comparing the $|\mathcal{S}|$ candidate batches, which has a time complexity of $\mathcal{O}(|\mathcal{S}|)$.

5. Performing interventions and dataset augmentation. Performing a batch of interventions and augmenting the interventional dataset are relatively simple operations and negligible to the runtime complexity.

6. Identifying Pareto-optimal solutions. Identifying Pareto-optimal solutions in within the interventional dataset $\mathcal{D}_N^I$ has complexity $\mathcal{O}(N_s^2 d_s)$ for each local MO-CBO problem w.r.t. $\mathbf{X}_s$, where d_s denotes the input dimension, and N_s the interventional data in $\mathcal{D}_N^I$ w.r.t. $\mathbf{X}_s$.

6 Experiments and results

This chapter aims to experimentally validate the proposed MO-CBO algorithm, CAUSAL PARETOSELECT, by testing its performance on synthetic graphs as well as evaluating its applicability for a real-world, domain-specific problem. The intention is to provide a comprehensive examination of the algorithm's effectiveness both in idealised synthetic settings and practical applications.

Firstly, section 6.1 focuses on the performance evaluation of CAUSAL PARETOSELECT through a series of experiments on synthetically generated structural causal models (SCMs). A benchmark study will compare CAUSAL PARETOSELECT to the naive approach for solving MO-CBO, which directly applies a MOBO algorithm on all treatment variables simultaneously. More specifically, these experiments are designed to test the algorithm across a range of performance metrics, such as cost-efficiency and mean prediction error towards the ground-truth causal Pareto front. We consider several graph configurations with different dependency structures including scenarios with and without unobserved confounders. Subsequently, in section 6.2, the focus shifts toward testing CAUSAL PARETOSELECT's functionality in a domain-specific context, using a causal model from a real-world example in healthcare. In contrast to the synthetic settings, this introduces a higher level of complexity, characterised by noisy function evaluations that pose a greater challenge to optimisation methods (Daulton et al., 2021). This experiment demonstrates the algorithm's potential for practical deployment in decision-making problems within multi-objective causal systems.

6.1 Benchmarking MO-CBO on synthetic problems

CAUSAL PARETOSELECT will be evaluated on three synthetic problems that feature different causal relationships and include graphs both with and without unobserved confounders. We benchmark our method by applying the standard multi-objective optimisation algorithm DGEMO (Konakovic Lukovic et al., 2020) to all treatment variables.

6.1.1 Performance indicators

In multi-objective optimisation, we require metrics to assess the quality of a given Pareto front approximation. These metrics evaluate both the convergence of the approximated front to the true front and the diversity of the solutions across the performance space. For a given optimisation problem, let $\mathbf{A}$ be the set of points from an approximated Pareto front. If the ground-truth Pareto front is known, it is possible to evaluate how well $\mathbf{A}$ approximates it given the following two metrics. Let $\mathbf{Z}$ be the set of points on the true Pareto front.

Generational distance (GD). A common performance indicator for evaluating a given Pareto front approximation is the so called *generational distance* (Schutze et al., 2012). It is the average distance from any point $\mathbf{a}_i \in \mathbf{A}$ to its closest point in the Pareto front $\mathbf{Z}$. Formally,

$$\mathrm{GD}(\mathbf{A}) = \left(\frac{1}{|\mathbf{A}|} \sum_{i=1}^{|\mathbf{A}|} d_i^p \right)^{1/p}, \tag{6.1}$$

for $p \in \mathbb{R}_{\geqslant 1}$, and d_i^p represents the Euclidean distance (e.g. $p = 2$) from $\mathbf{a}_i$ to its nearest point in $\mathbf{Z}$.

Inverted generational distance (IGD). The so called *inverted generational distance* measures the distance from any point $\mathbf{z}_i \in \mathbf{Z}$ to its closest point in $\mathbf{A}$ (Schutze et al., 2012). Thereby, it can serve as an indicator for the coverage provided by the approximated front. Formally,

$$\mathrm{IGD}(\mathbf{A}) = \left(\frac{1}{|\mathbf{Z}|} \sum_{i=1}^{|\mathbf{Z}|} \hat{d}_i^{\,p} \right)^{1/p}, \tag{6.2}$$

for $p \in \mathbb{R}_{\geqslant 1}$, and $\hat{d}_i^{\,p}$ represents the Euclidean distance (e.g. $p = 2$) from $\mathbf{z}_i$ to its nearest point in $\mathbf{A}$.

Since Causal ParetoSelect aims to discover the causal Pareto front, both the generational distance and the inverted generational distance will be used as indicators of the algorithm's performance, setting $p = 2$. The smaller both distances are, the better Causal ParetoSelect performs in approximating and discovering the causal Pareto front. These metrics provide a combined evaluation of both the convergence and coverage properties of the algorithm's output.

6.1.2 Experimental setup

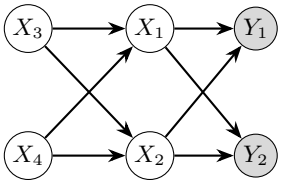
Baseline. The baseline method is given by the multi-objective optimisation algorithm dgemo (Konakovic Lukovic et al., 2020), applying it on all treatment variables simultaneously. Therefore, the baseline is defined by the optimisation problem with objective functions $\mu_i(\mathbf{X}, \cdot) : \mathcal{D}(\mathbf{X}) \to \mathbb{R}$, $\mathbf{x} \mapsto \mathbb{E}[Y_i|\mathrm{do}(\mathbf{X} = \mathbf{x})]$, for $i = 1, \ldots, m$, neglecting available causal information. Since mo-cbo is a new problem formulation, there are no other sample-efficient benchmark paradigms available, that can operate with only a few hundred samples.

Setup. In each experiment, we set an upper bound for the total interventional cost, called *interventional budget*. The cost of any intervention $\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)$ is fixed at 1, meaning that the interventional budget directly limits the number of interventions that can be performed. While we use this fixed cost structure, it is also possible to define alternative cost functions, where the cost of an intervention depends on $\mathbf{X}_s$ and/or $\mathbf{x}_s$. Moreover, given an exploration set $\mathcal{S} \subseteq \mathcal{P}(\mathbf{X})$, we assume the availability of an interventional dataset $\mathcal{D}^I = \{((\mathbf{X}_s, \mathbf{x}_s^k), \boldsymbol{\mu}(\mathbf{X}_s, \mathbf{x}_s^k))\}_{k=1,s=1}^{K,|\mathcal{S}|}$. Both the baseline algorithm and Causal ParetoSelect are executed until the interventional budget is exhausted. For reproducibility, all experiments are run across 10 random seeds, and

we report the average performance metrics. Note that each seed results in a different initialisation of the interventional data $\mathcal{D}^I$. Additionally, the batch size is set to 10. Prior to conducting each synthetic experiment, we compute the ground-truth causal Pareto front $\mathcal{P}_f^{\mathcal{C}}(\mathcal{P}(\mathbf{X}))$ for benchmark purposes.

6.1.3 Synthetic-1

We introduce the first synthetic MO-CBO problem in our experimental study, referred to as Synthetic-1, which is defined by the causal graph $\mathcal{G}$ and associated structural assignments presented in Figure 6.1. Here, the set of treatment variables is $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$, and the set of output variables is $\mathbf{Y} = \{Y_1, Y_2\}$. The interventional domains are specified as $\mathcal{D}(X_1), \mathcal{D}(X_2) = [-1, 3]$ and $\mathcal{D}(X_3), \mathcal{D}(X_4) = [-1, 1]$. Moreover, all exogenous variables follow the standard normal distribution, and there are no unobserved confounders.



$$X_1 = \exp((X_3 - X_4)/2) + U_{X_1} \qquad Y_1 = (X_1 + X_2)^2 + U_{Y_1}$$
$$X_2 = ((X_3 - X_4)/2)^3 + U_{X_2} \qquad Y_2 = (X_1 + X_2 - 10)^2 + U_{Y_2}$$
$$X_3 = U_{X_3}$$
$$X_4 = U_{X_4} \qquad U_{X_i}, U_{Y_i} \sim \mathcal{N}(0, 1)$$

Figure 6.1: Synthetic-1. An SCM consisting of four treatment and two output variables (shaded nodes), with no unobserved confounders.

The minimal intervention sets are $\mathbb{M}_{\mathcal{G},\mathbf{Y}} = \{\varnothing, \{X_1\}, \{X_2\}, \{X_3\}, \{X_4\}, \{X_1, X_2\}, \{X_1, X_3\}, \{X_1, X_4\}, \{X_2, X_3\}, \{X_2, X_4\}, \{X_3, X_4\}, \{X_1, X_3, X_4\}, \{X_2, X_3, X_4\}\}$. Since there are no unobserved confounders present, there is only one possibly-optimal minimal intervention set, namely $\mathbb{P}_{\mathcal{G},\mathbf{Y}} = \{\text{pa}(\mathbf{Y})_{\mathcal{G}}\} = \{\{X_1, X_2\}\}$.

We now compare Causal ParetoSelect to the baseline method in regards to their accuracy in approximating the causal Pareto front for Synthetic-1. All experiments are executed until the specified interventional budget of 800 cost units is exhausted.

Performance results. The performance metrics from the experiments on Synthetic-1 are displayed in Table 6.1. Note that CPS+$\mathbb{M}_{\mathcal{G},\mathbf{Y}}$ exhibits the worst performance across both metrics, which is to be expected given that this method explores a large number of intervention

Method	GD	IGD
Baseline	**0.1908**	1.4446
CPS+$\mathbb{M}_{\mathcal{G},\mathbf{Y}}$	0.2135	1.9753
CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$	0.2152	**0.7121**

Table 6.1: Synthetic-1. Performance metrics from the baseline experiment, CPS+$\mathbb{M}_{\mathcal{G},\mathbf{Y}}$, and CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ after exhausting the interventional budget. Results are averaged across initialisations with multiple seeds.

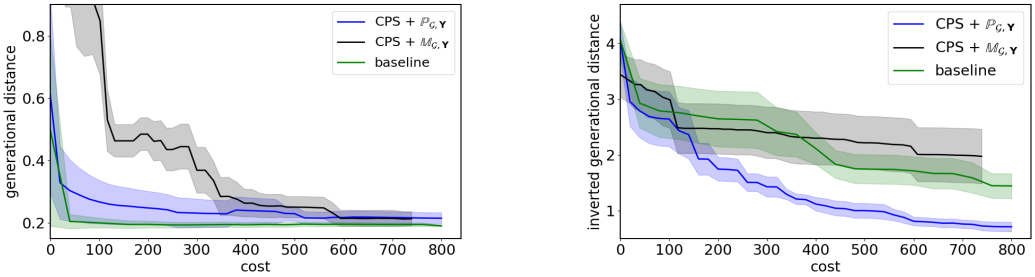


Figure 6.2: SYNTHETIC-1. Convergence to the causal Pareto front across all experiments, measured by the generational distance (left) and inverted generational distance (right). Shaded areas represent ± standard deviation from initialisations with multiple seeds.

sets with a very limited budget. Moreover, the differences in generational distance are marginal across all experiments while the inverted generational distance is significantly reduced by CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$, indicating that diverse regions of the causal Pareto front can be discovered with our method as compared to the baseline algorithm.

We proceed by providing a more comprehensive analysis and consider the progression of each performance metric relative to the total interventional cost. The corresponding results are displayed in Figure 6.2. Notably, CPS+$\mathbb{M}_{\mathcal{G},\mathbf{Y}}$ exhibits significantly slower improvements in comparison to the other methods. This is indeed to be expected due to the Pareto front exploration within the set of minimal intervention sets, most of which cannot yield dominant trade-offs. Moreover, the baseline method improves faster in terms of the generational distance than CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$, albeit the difference is marginal. However, a lower inverted generational distance can be reached with CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$, yielding a more diverse causal Pareto front approximation at lower cost. In summary, CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ balances performance progression and cost efficiency, achieving outcomes that are on par with or better than the baseline approach, while also reducing interventional cost. We analyse these results for each experiment separately:

Baseline

Recall that the baseline algorithm involves manipulating all treatment variables simultaneously, and therefore corresponds to an optimisation problem defined by the objective functions $\mu_i(\mathbf{X}, \cdot) : \mathcal{D}(\mathbf{X}) \to \mathbb{R}, \mathbf{x} \mapsto \mathbb{E}[Y_i|\text{do}(\mathbf{X} = \mathbf{x})]$, for $i = 1, 2$. Since $\mathcal{D}(\mathbf{X}) \subseteq \mathbb{R}^4$, this corresponds to a typical multi-objective optimisation problem. Again, intervening on all variables simultaneously is a brute-force strategy that does not leverage known causal relationships. As depicted in Figure 6.3 when X_1 and X_2 are already being intervened upon, a further manipulation of X_3 and X_4 is redundant as the effects will not be propagated to the targets. In other words, $\mathbf{X}$ is not a minimal intervention set.

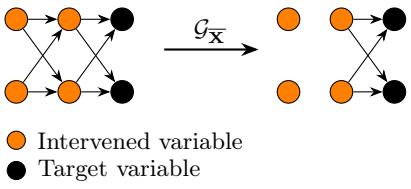


Figure 6.3: SYNTHETIC-1. Intervening on $\mathbf{X}$.

The causal Pareto front approximation from the baseline experiment is presented in Figure 6.4. We observe that parts of the causal Pareto front are being well approximated and the identified Pareto-optimal solutions deviate only marginally from the ground truth, visually confirming the result from Table 6.1. This slight deviation is likely caused by the noise attributed to the exogenous variables U_{Y_i}, $i = 1, 2$. Indeed, we argue that a negligible generational distance is to be expected given the absence of unobserved confounders: $\text{pa}(\mathbf{Y})_{\mathcal{G}} = \{X_1, X_2\}$ is the only possibly-optimal minimal intervention set, and $\{X_1, X_2\} \subseteq \mathbf{X}$ ensures that dominant trade-offs between Y_1 and Y_2 can be identified through interventions on $\mathbf{X}$. However, intervening on additional variables, X_3 and X_4, incurs unnecessary costs, highlighting the inefficiency of the baseline strategy. The absence of solutions in the region $Y_1 \sim [23, 35]$ stems from the interventional budget limiting additional iterations, in which DGEMO can explore and identify diverse regions of the Pareto front.

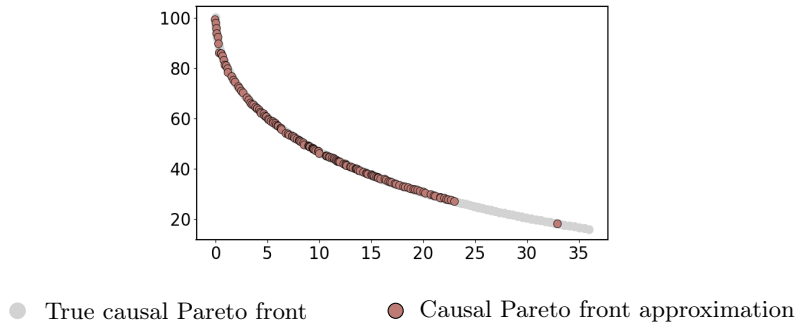
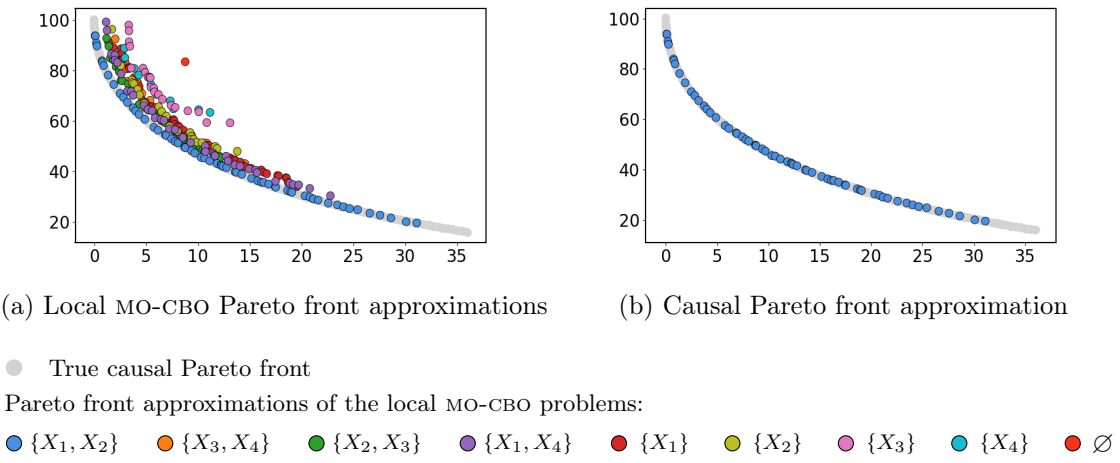


Figure 6.4: SYNTHETIC-1. Causal Pareto front approximation from the baseline experiment.

CPS+$\mathbb{M}_{\mathcal{G},\mathbf{Y}}$: CAUSAL PARETOSELECT with minimal intervention sets

We address SYNTHETIC-1 with CAUSAL PARETOSELECT, solving the local MO-CBO problems associated to the minimal intervention sets. In each iteration, the algorithm chooses the set with the highest relative hypervolume improvement to intervene upon. The visual results from this experiment are displayed in Figure 6.5. Specifically, Figure 6.5a illustrates the approximated Pareto fronts of each local MO-CBO problem, while Figure 6.5b isolates solutions that are Pareto-optimal. We observe that certain intervention sets significantly outperform others in minimising the target variables: For instance, interventions on $\text{pa}(\mathbf{Y})_{\mathcal{G}}$ consistently yield causally Pareto-optimal solutions, whereas those on $\{X_3\}$ produce the least favourable Pareto front. Indeed, all causally Pareto-optimal solutions discovered in this experiment originate from interventions on $\text{pa}(\mathbf{Y})_{\mathcal{G}}$, visually underscoring that it is the sole possibly-optimal minimal intervention set. Compared to both the baseline scenario (and experiments involving the possibly-optimal set), the causal Pareto front approximation appears significantly more sparse, as also indicated by the inverted generational distance in Table 6.1. This sparsity is attributed to the algorithm's alternating selection of different intervention sets, leading to fewer iterations where interventions on $\text{pa}(\mathbf{Y})_{\mathcal{G}}$ are evaluated and incorporated into the causal Pareto front.



(a) Local MO-CBO Pareto front approximations

(b) Causal Pareto front approximation

Figure 6.5: Synthetic-1. Experimental results from Causal ParetoSelect with minimal intervention sets. Not all minimal intervention sets were chosen for evaluation.

CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$: Causal ParetoSelect with possibly-optimal sets

Finally, we address Synthetic-1 with Causal ParetoSelect focusing on the possibly-optimal minimal intervention set, $\text{pa}(\mathbf{Y})_{\mathcal{G}}$. An illustration of the set in both the original and intervened graph is shown in Figure 6.6. Notably, intervening on $\text{pa}(\mathbf{Y})_{\mathcal{G}}$ achieves the same causal effect on $\mathbf{Y}$ as the baseline approach, while avoiding unnecessary interventions the remaining variables, X_3 and X_4. The causal Pareto front approximation from this experiment is presented in Figure 6.7. We observe that all identified solutions deviate only marginally from the ground

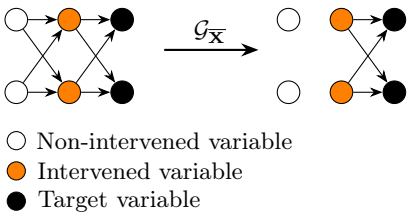


Figure 6.6: Synthetic-1. Intervening on $\text{pa}(\mathbf{Y})_{\mathcal{G}}$.

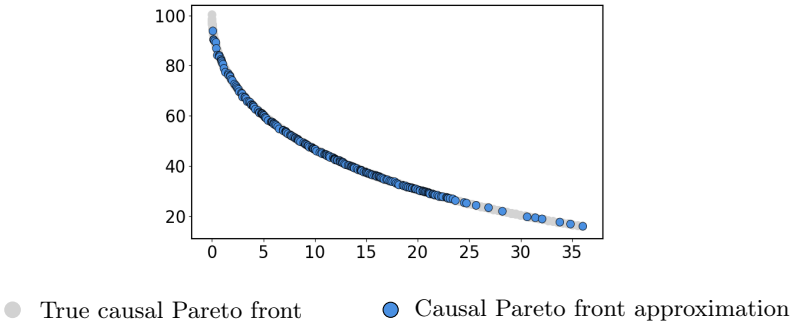
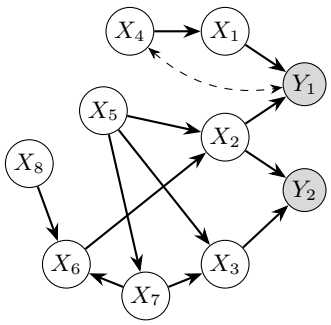


Figure 6.7: Synthetic-1. Experimental Results for Causal ParetoSelect with possibly-optimal minimal intervention sets.

truth, as also indicated by the low generational distance in Table 6.1. Again, this expected and caused by the noise attributed to the exogenous variables U_{Y_i}, $i = 1, 2$. Notably, CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ significantly outperforms CPS+$\mathbb{M}_{\mathcal{G},\mathbf{Y}}$ in both metrics, which is to be expected and has already been discussed. Furthermore, CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ also significantly improves the coverage of the approximated front as compared to the baseline approach. By avoiding unnecessary interventions on X_3 and X_4, it reduces overall costs, allowing for more exploratory interventions on $\text{pa}(\mathbf{Y})_{\mathcal{G}}$ within the given interventional budget. This property is reflected by the the progression of the inverted generational in Figure 6.2, as well as the final value of the inverted generational distance given Table 6.1.

6.1.4 Synthetic-2

Synthetic-2 is the next mo-cbo problem of this experimental study, defined by the causal graph $\mathcal{G}$ and associated structural equations in Figure 6.8. Here, the set of treatment variables is $\mathbf{X} = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8\}$ and the set of output variables is $\mathbf{Y} = \{Y_1, Y_2\}$. The interventional domains are $\mathcal{D}(X_1) = [-2, 5]$, $\mathcal{D}(X_4) = [-4, 5]$ and $\mathcal{D}(X_i) = [0, 5]$ for all $i = 1, \ldots, 8$ with $i \neq 1, 4$. Moreover, the exogenous variables U_{X_i}, U_{Y_i} follow a Gaussian distribution, and there is an unobserved confounder U influencing the target variable Y_1 and its ancestor X_4.



$$
\begin{aligned}
X_1 &= X_4/2 + 0.1U_{X_1} \\
X_2 &= \exp(X_5 + X_6 - 10) + U_{X_2} \\
X_3 &= \ln(1 + (X_5 + X_7)/10) + U_{X_3} \\
X_4 &= U + (U_{X_4}/20)^2 \\
X_5 &= U_{X_5} \\
X_6 &= \exp(X_5 + X_7 + X_8 - 30) + U_{X_6} \\
X_7 &= U_{X_7} \\
X_8 &= U_{X_8}
\end{aligned}
$$

$$
\begin{aligned}
Y_1 &= \ln(1 + X_1^2) + 2 \cdot X_2^2 - X_1 \cdot X_2 \cdot U/2 + U_{Y_1}^2 \\
Y_2 &= \sin(X_2^2) - X_3^2 - X_2 \cdot X_3 + 50 + U_{Y_2}^2
\end{aligned}
$$

$$
\begin{aligned}
&U \in \{-4, 4\},\ P(U = -4) = P(U = 4) = 0.5 \\
&U_{X_i}, U_{Y_i} \sim \mathcal{N}(0, 0.1)
\end{aligned}
$$

Figure 6.8: Synthetic-2. An scm consisting of eight treatment and two output variables, including an unobserved confounder (dashed bi-directed edge) affecting one output and its ancestor.

In this subsection, we compare Causal ParetoSelect to the baseline method in regards to their performance in approximating the causal Pareto front for Synthetic-2. However, we omit the experiment involving minimal intervention sets, i.e. CPS+$\mathbb{M}_{\mathcal{G},\mathbf{Y}}$, since it is an ineffective strategy, as discussed in the last subsection. Due to the presence of an unob-

served confounder, there are two possibly-optimal minimal intervention sets, namely $\mathbb{P}_{\mathcal{G},\mathbf{Y}} = \{\{X_2, X_3\}, \{X_1, X_2, X_3\}\}$. Again, all experiments are executed until the specified interventional budget of 800 cost units is exhausted.

Performance results. The resulting performance metrics from experiments on SYNTHETIC-2 are displayed in Table 6.2. In summary, our algorithm CAUSAL PARETOSELECT significantly outperforms the baseline method in regards to approximating the causal Pareto front, as measured by the generational and inverted generational distance. However, it is important to note here that the magnitude of this deviation strongly depends on the given MO-CBO problem, especially when unobserved confounders are present.

Method	GD	IGD
Baseline	5.8231	4.2724
CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$	**1.7389**	**0.3077**

Table 6.2: SYNTHETIC-2. Performance metrics from the baseline experiment and CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$, after exhausting the interventional budget. Results are averaged across initialisations with multiple seeds.

We continue by providing a more detailed analysis and consider the progression of each performance metric relative to the total interventional cost throughout the experiments. The corresponding results are displayed in Figure 6.9. CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ outperforms the baseline approach and is able to reach significantly lower values of the generational and inverted general distance. It does so at reduced interventional costs. In other words, the metrics indicate that CAUSAL PARETOSELECT provides are more accurate approximation of the causal Pareto front, while requiring fewer interventions.

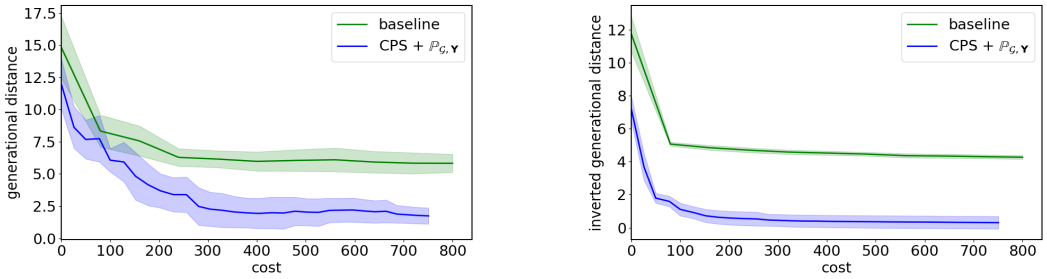


Figure 6.9: SYNTHETIC-2. Convergence to the causal Pareto front across all experiments, measured by the generational distance (left) and inverted generational distance (right). Shaded areas represent ± standard deviation from initialisations with multiple seeds.

Baseline

The baseline algorithm involves manipulating all treatment variables simultaneously. However, the effects of variables in $\{X_4, X_5, X_6, X_7, X_8\}$ will not be propagated to the targets,

incurring unnecessary cost. Moreover, this strategy breaks the causal path $X_4 \to X_1 \to Y_1$, letting the unobserved confounder affect Y_1 without its effects being also propagated through the aforementioned path, see Figure 6.10.

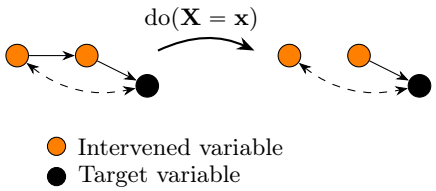


Figure 6.10: Synthetic-2. UC-territory after intervening on $\mathbf{X}$.

The visual results from this experiment are depicted in Figure 6.11. We observe that Pareto front approximation is very noisy which is attributed to the the unobserved confounder causing highly fluctuating function evaluations at random. More importantly, the baseline method is unable to approximate the true causal Pareto front. The identified solutions are not Pareto-optimal and lie far beyond the causal Pareto front. This observation is represented by the values of the generational distance and inverted generational distance, indicating that the approximation and the ground truth deviate significantly.

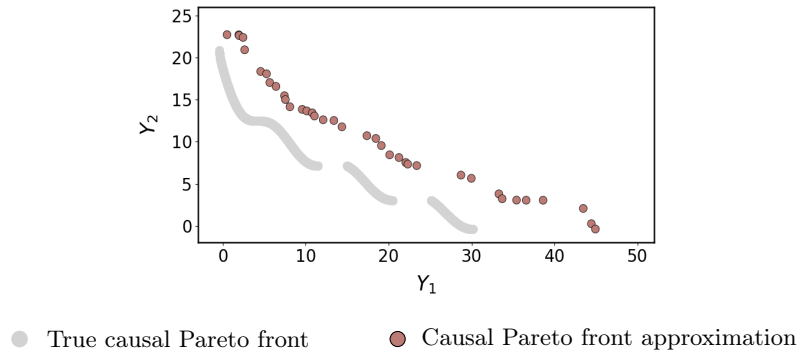
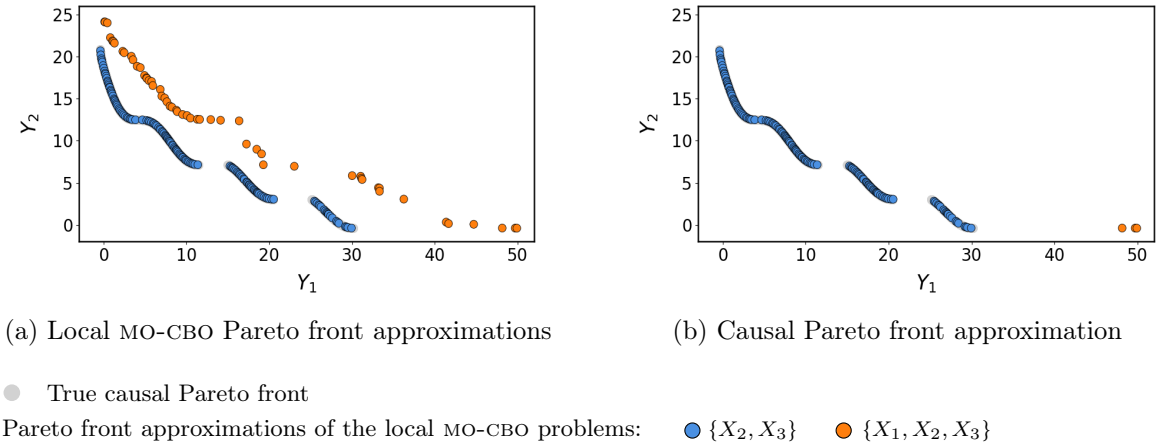


Figure 6.11: Synthetic-2. Causal Pareto front approximation from the baseline experiment.

CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$: Causal ParetoSelect with possibly-optimal sets

We address Synthetic-2 with Causal ParetoSelect focusing on the set of all possibly-optimal minimal intervention sets, i.e. $\mathbb{P}_{\mathcal{G},\mathbf{Y}} = \{\{X_2, X_3\}, \{X_1, X_2, X_3\}\}$. Albeit the causal effect of $\{X_1, X_2, X_3\}$ on $\mathbf{Y}$ is known from the baseline experiment, it is unclear whether CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ improves at solving the its local mo-cbo problem, given that the design space is reduced from eight to three dimensions. Conversely, intervening on $\{X_2, X_3\}$ allows the effect of the unobserved confounder to be propagated through the path $X_4 \to X_1 \to Y_1$.

The visual results of this experiment are presented in Figure 6.12. We observe that the Pareto front approximation from the local mo-cbo problem w.r.t. $\{X_1, X_2, X_3\}$ is noisy and does not approximate the ground truth well. However, $\{X_2, X_3\}$ fully covers the causal Pareto front, which was already indicated by the negligible inverted generational distance in Table 6.2. We explain why this result occurs: As already mentioned, intervention on $\{X_2, X_3\}$ let the effect of U be propagated through the path $X_4 \to X_1 \to Y_1$, while U also effects Y_1 directly. The structural assignment of Y_1 includes the term $-X_1 \cdot X_2 \cdot U/2$. Not intervening on X_1, causes



(a) Local MO-CBO Pareto front approximations

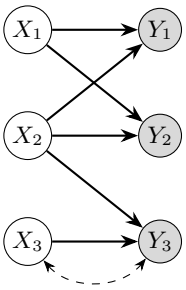
(b) Causal Pareto front approximation

Figure 6.12: Synthetic-2. Experimental results for Causal ParetoSelect with the possibly-optimal minimal intervention sets.

this term to always be negative, yielding lower function values for Y_1. However, if we do intervene on X_1, this term is positive with probability 0.5, causing higher values for Y_1 in the averaged outcomes. On another note, noisy function evaluations result in the inclusion of few solutions from the local MO-CBO problem w.r.t. $\{X_1, X_2, X_3\}$ into the causal Pareto front approximation. This artifact contributes to an increased generational distance, as seen in Table 6.2. Nonetheless, our method Causal ParetoSelect is able to accurately address Syntehtic-2, while the baseline method fails to incorporate causal information and this yields suboptimal outcomes.

6.1.5 Synthetic-3

We propose the last synthetic MO-CBO problem of our experimental study, called Synthetic-3, that necessitates simultaneous optimisation of three output variables. It is defined by the causal graph $\mathcal{G}$ and associated structural assignments presented in Figure 6.13.



$$X_1 = U_{X_1}$$
$$X_2 = U_{X_2}$$
$$X_3 = U + U_{X_3}$$

$$Y_1 = X_1^2 + X_2^2 + U_{Y_1}$$
$$Y_2 = (X_2 - 5)^2 + (X_1 - 10)^2 + U_{Y_2}$$
$$Y_3 = (X_2 - 10)^2 - X_3 \cdot U + U_{Y_3}$$

$$U \in \{-5, 5\},\ P(U = -5) = P(U = 5) = 0.5$$
$$U_{X_i}, U_{Y_i} \sim \mathcal{N}(0, 0.5)$$

Figure 6.13: Synthetic-3. An SCM consisting of three treatment and three target variables, and an unobserved confounder (dashed bi-directed edge) affecting one output and its parent.

Here, the set of treatment variables is $\mathbf{X} = \{X_1, X_2, X_3\}$, and the set of output variables is $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$. The interventional domains are given as $\mathcal{D}(X_1), \mathcal{D}(X_2) = [0, 5]$ and $\mathcal{D}(X_3) = [-5, 5]$. Moreover, all exogenous variables follow a Gaussian distribution, and there is an unobserved confounder U influencing the target variable Y_3 and its parent X_3.

In this subsection, we compare Causal ParetoSelect to the baseline method in regards to their performance in approximating the causal Pareto front for Synthetic-3. Due to the presence of an unobserved confounder, there are two possibly-optimal minimal intervention sets, given by $\mathbb{P}_{\mathcal{G},\mathbf{Y}} = \{\{X_1, X_2\}, \{X_1, X_2, X_3\}\}$. Again, all experiments are executed until the specified interventional budget of 800 cost units is exhausted.

Performance results. The resulting performance metrics from experiments on Synthetic-3 are displayed in Table 6.3. In summary, Causal ParetoSelect outperforms the baseline method in approaching the causal Pareto front, as evidenced by lower generational and inverted generational distances. Again, it is important to emphasise that the extent of this improvement depends on the specific mo-cbo problem, particularly on the structural equations involving unobserved confounders. Nonetheless, it is surprising that our method yields such high deviations from the ground truth. We analyse this further in the upcoming paragraphs.

Method	GD	IGD
Baseline	11.9295	14.7913
CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$	**7.1329**	**4.92478**

Table 6.3: Synthetic-3. Performance metrics from the baseline experiment and CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$, after exhausting the interventional budget. Results are averaged across initialisations with multiple seeds.

We proceed with examining how each performance metric evolves in relation to the total interventional cost over the course of the experiments. The results are illustrated in Figure 6.14.

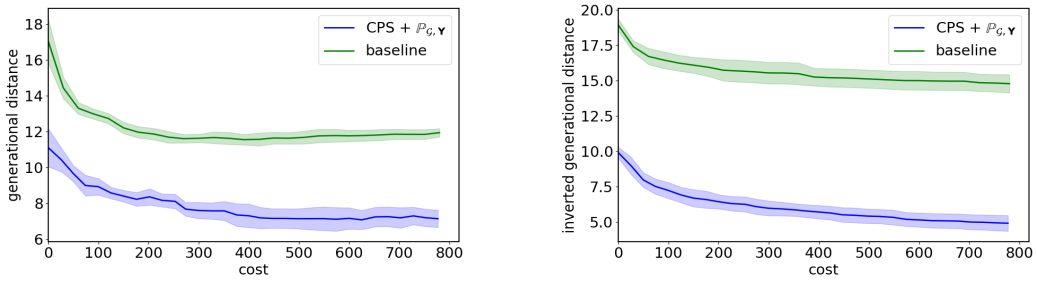


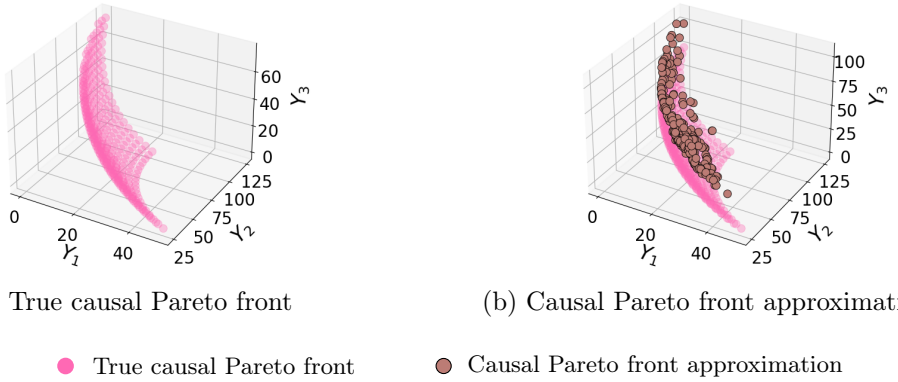
Figure 6.14: Synthetic-3. Convergence to the causal Pareto front across all experiments, measured by the generational distance (left) and inverted generational distance (right). Shaded areas represent ± standard deviation from initialisations with multiple seeds.

CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ surpasses the baseline method, achieving better values for both the generational distance and the inverted generational distance, at reduced costs. The results suggest that Causal ParetoSelect identifies a more precise approximation of the causal Pareto front, while performing fewer interventions.

Baseline

Since the baseline algorithm involves manipulating all treatment variables simultaneously, it allows the unobserved confounder U to affect Y_3 without its effects being propagated through the path $X_3 \to Y_3$. This leads to suboptimal outcomes due to the structural equation for Y_3, where $-X_3 \cdot U$ is always negative if X_3 is not intervened upon.

The visual results of the baseline experiment on Synthetic-3 are depicted in Figure 6.15. Indeed, we notice that the Pareto front approximation is shifted upwards in the direction of Y_3. We conclude that the baseline method leads to suboptimal solutions, and further analysis shows that it does not yield Pareto-optimal solutions. This observation is also represented by the generational distance and inverted generational distance in Table 6.3, indicating that the approximation and the ground truth deviate significantly.



(a) True causal Pareto front

(b) Causal Pareto front approximation

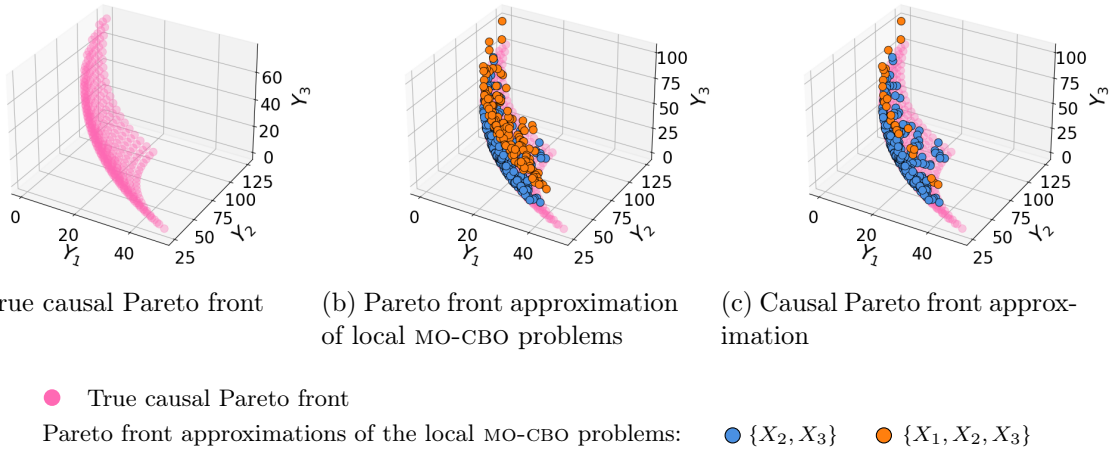
Figure 6.15: Synthetic-3. Causal Pareto front approximation from the baseline experiment.

CPS+$\mathbb{P}_{\mathcal{G},\mathbf{Y}}$: Causal ParetoSelect with possibly-optimal sets

Finally, we address Synthetic-3 with Causal ParetoSelect focusing on the set of possibly-optimal minimal intervention sets, i.e. $\mathbb{P}_{\mathcal{G},\mathbf{Y}} = \{\{X_2, X_3\}, \{X_1, X_2, X_3\}\}$. Albeit the causal effect of $\{X_1, X_2, X_3\}$ on $\mathbf{Y}$ is known from the baseline, the precise causal effect of $\{X_1, X_2\}$ on the outcomes is unknown.

The results of this experiment are presented in Figure 6.16. We observe that the Pareto front approximation for the local mo-cbo problem w.r.t. $\{X_1, X_2\}$ closely aligns with the true causal Pareto. A further analysis shows that the high generational and inverted generational distances reported in Table 6.3 are due to noisy function evaluations. This leads to the inclusion of a few solutions from the local mo-cbo problem with respect to $\{X_1, X_2, X_3\}$ into the causal Pareto front approximation. These solutions deviate significantly from the ground truth, explaining

the degraded performance metrics in Table 6.3.



(a) True causal Pareto front

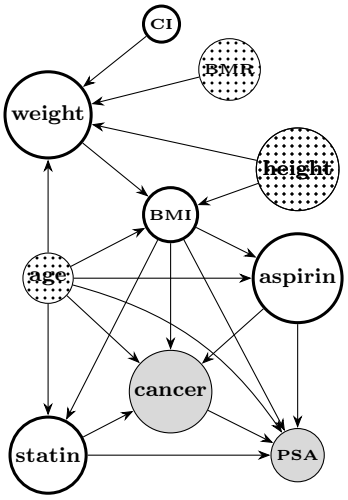
(b) Pareto front approximation of local MO-CBO problems

(c) Causal Pareto front approximation

Figure 6.16: SYNTHETIC-3. Experimental results for CAUSAL PARETOSELECT with the possibly-optimal minimal intervention sets.

6.2 Validating MO-CBO on a healthcare application

In this section, we transition from idealised synthetic settings to a structural causal model grounded in a real-world healthcare application. More specifically, we analyse the MO-CBO problem defined by the causal graph and structural equations in Figure 6.17, and refer to it as HEALTH. This model originates from previous works of Ferro et al., 2015, and is based on real-world causal relationships. It captures the interactions between the variables BMI, age, aspirin and statin usage (among others), and their effects on prostate-specific-antigen (PSA) levels and prostate cancer occurrence. The variable aspirin indicates the daily aspirin regimen, while statin denotes whether the subject is taking any statin medication. Additionally, PSA represents the total antigen level circulating in a subject's blood, measured in ng/mL. Originally, PSA is the sole target variable as it is an easily measurable indicator for prostate cancer. However, since it is not guaranteed to be a fully reliable measure, it may be critical to also take into account the actual cancer occurrence. To this end, we reformulate to a MO-CBO problem by treating both cancer occurrence and PSA levels as target variables. The treatment variables shall include BMI, aspirin and statin usage with interventional domains $\mathcal{D}(\text{BMI}) = [20, 30]$ and $\mathcal{D}(\text{aspirin}), \mathcal{D}(\text{statin}) = [0, 1]$, while age is set to be non-manipulative. The aim is to reduce cancer occurrence and PSA levels simultaneously, identifying non-dominated trade-offs between these objectives. In the previous literature (e.g. Gultchin et al., 2023), age was modelled as uniformly distributed between 55 and and 75; however we can choose to focus on specific age groups of interest. U_{age} is now distributed as a Gaussian random variable with mean 65 and standard deviation 1, focusing on individuals close to the age of 65. Moreover, there is a single possibly-optimal minimal intervention set, $\mathbb{P}_{\mathcal{G},\mathbf{Y}} = \{\{\text{BMI}, \text{aspirin}, \text{statin}\}\}$, which consists of



$$
\begin{aligned}
\text{CI} &= U_{\text{CI}}, \quad U_{\text{CI}} \sim \mathcal{U}(-100, 100) \\
\text{BMR} &= 1500 + 10U_{\text{BMR}}, \quad U_{\text{height}} \sim t\mathcal{N}(-1, 2) \\
\text{height} &= 175 + 10U_{\text{height}}, \quad U_{\text{height}} \sim t\mathcal{N}(-0.5, 0.5) \\
\text{age} &= U_{\text{age}}, \quad U_{\text{age}} \sim \mathcal{N}(65, 1) \\
\text{weight} &= (\text{BMR} + 6.8\text{age} - 5\text{height})/(13.7 + \text{CI}150/7716) \\
\text{BMI} &= \text{weight}/(\text{height}/100)^2 \\
\text{aspirin} &= \sigma(-8 + 0.1\text{age} + 0.03\text{BMI}), \quad \sigma(x) = \tfrac{1}{1+e^{-x}} \\
\text{statin} &= \sigma(-13 + 0.10\text{age} + 0.20\text{BMI}) \\
\text{cancer} &= \sigma(2.2 - 0.05\text{age} + 0.01\text{BMI} - 0.04\text{statin} + 0.02\text{aspirin}) \\
\text{PSA} &= 6.8 + 0.04\text{age} - 0.15\text{BMI} - 0.6\text{statin} + 0.55\text{aspirin} + \text{cancer} + U_{\text{PSA}} \\
& \qquad\qquad U_{\text{PSA}} \sim \mathcal{N}(0, 0.4)
\end{aligned}
$$

Figure 6.17: HEALTH. An SCM with relations between variables such as age, BMI, aspirin and statin usage, and their effects on PSA levels and prostate cancer risk (Gultchin et al., 2023). $\mathcal{U}(\cdot, \cdot)$ denotes a uniform distribution and $t\mathcal{N}(a, b)$ a standard Gaussian distribution truncated between a and b. Dotted nodes are non-manipulative, thick edges represent treatment variables and the target variables are given as shaded nodes.

the manipulative parents of the target variables. The single-objective version of HEALTH has previously been used to demonstrate the applicability of CBO (Aglietti et al., 2020), as well as for several of its variants (e.g. Gultchin et al., 2023 and Aglietti et al., 2023). In this section, we aim to validate our approach in a similar manner by demonstrating its practical applicability. Rather than benchmarking our method, the focus is on showcasing its feasibility. We apply CAUSAL PARETOSELECT to the MO-CBO problem HEALTH, using $\mathbb{P}_{\mathcal{G},\mathbf{Y}}$ as the exploration set. We set an interventional budget of 90 cost units, limiting our algorithm to three iterations. The Pareto front approximation of this experiment is depicted in Figure 6.18. We observe that the optimal trade-offs between cancer and PSA levels are linear and CAUSAL PARETOSELECT can

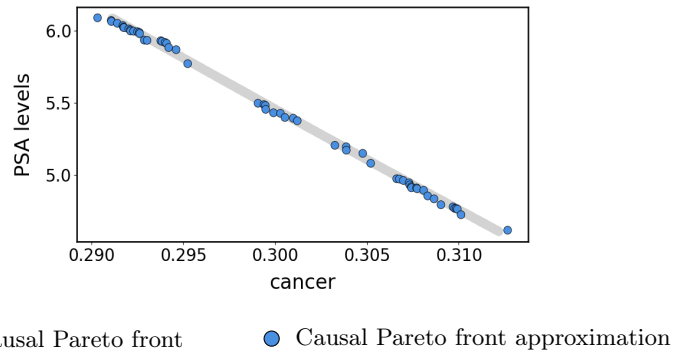


Figure 6.18: HEALTH. Experimental results for CAUSAL PARETOSELECT with the possibly-optimal minimal intervention set.

indeed detect this relationship, even though the SCM causes noisy outcomes. The true causal Pareto set is given by $\mathcal{P}_s^{\mathcal{C}}(\mathcal{P}(\mathbf{X})) = \{(\{\text{BMI}, \text{aspirin}, \text{statin}\}, (x, 0.0, 1.0) \mid x \in [20, 30]\})\}$, where $\mathbf{X}$ denotes the set of all variables that can be intervened upon. CAUSAL PARETOSELECT does correctly identify solutions up to small perturbations, which is confirmed by the generational distance being precisely 0.0019. The inverted generational distance is higher, and lies at 0.02968, indicating that parts of the causal Pareto front were not identified. This aligns with the visual results from Figure 6.18. However, CAUSAL CARETOSELECT can give insights on such real-world processes: The approximated front reveals that, regardless of the BMI, the optimal drug dosages for reducing both cancer and PSA risk remain constant. However, the BMI still plays a significant role in influencing cancer and PSA levels, as demonstrated by the variation in the trade-offs in Figure 6.18. Specifically, different BMI result in different outcomes for the balance between cancer occurrence and PSA levels.

7 Discussion

In this chapter, we discuss our methodology and results towards *multi-objective causal Bayesian optimisation* (MO-CBO), the new problem class introduced in this thesis. A MO-CBO problem involves sequentially performing interventions in a causal system in order to optimise multiple outcome variables simultaneously, subject to a budgetary constraint as interventions are assumed to be costly. We have proposed an algorithm, called Causal ParetoSelect, that addresses MO-CBO by iteratively solving local MO-CBO problems based on relative hypervolume improvement. The selection of these local problems is guided by the theory developed in chapter 4, where we establish a framework for identifying possibly-optimal minimal intervention sets, building on the work of Lee and Bareinboim, 2018. This includes a non-trivial analysis of the graph structure to derive graphical criteria for selecting intervention sets feasible to consider for our optimisation objective. Notably, the goal of Causal ParetoSelect is to approximate the causal Pareto front by identifying Pareto-optimal intervention set-value pairs. Given a causal graph $\langle \mathcal{G}, \mathbf{X}, \mathbf{Y} \rangle$, both theoretical insights from chapter 4 as well as experimental results from chapter 6, demonstrate that there are two scenarios to be distinguished:

Case 1: There are no unobserved confounders between a target variable Y_i and any of its ancestors in $\text{an}(Y_i)_{\mathcal{G}}$, for all $i = 1, \ldots, m$.

Case 2: There exists at least one target variable Y_i, $1 \leqslant i \leqslant m$, which has an unobserved confounder with at least one of its ancestors in $\text{an}(Y_i)_{\mathcal{G}}$.

In case 1, there exists a single possibly-optimal minimal intervention set comprised of the outcomes' parents, i.e. $\text{pa}(Y_1, \ldots, Y_m)_{\mathcal{G}}$. In this case, both Causal ParetoSelect and traditional multi-objective Bayesian optimisation, can identify solutions on the causal Pareto front. This result is theoretically grounded from comparing intervention sets in chapter 4, and was thereafter experimentally illustrated in subsection 6.1.3. However, Causal ParetoSelect is more cost-efficient and finds a more diverse range of solutions, see again subsection 6.1.3. Conversely, in case 2, unobserved confounders may lead to the existence of several possibly-optimal minimal intervention sets. In this scenario, the causal Pareto front cannot always be found by traditional multi-objective Bayesian optimisation, rendering Causal ParetoSelect particularly useful. This discrepancy arises as unobserved confounders propagate their effects from $\text{an}(Y_i)_{\mathcal{G}}$ to Y_i, and, depending on the functional relationships between the variables, disrupting this propagation causes traditional methods to yield suboptimal solutions.

Next, we discuss the limitations of Causal ParetoSelect when solving MO-CBO problems.

Surrogate model. Since the outcome variables $Y_1, \ldots, Y_m$ can share endogenous confounders, we conjecture that modelling the objectives $\mathbb{E}[Y_i|\text{do}(\cdot)]$ as independent can lead to inefficiencies. Rather, integrating a more complex Gaussian process surrogate model where information is shared across objectives may increase information gain as well as cost-efficiency. To this end, one may consider multi-task Gaussian processes which are designed to learn a shared covariance function (Bonilla et al., 2007).

Non-manipulative variables. The methodology for selecting possibly-optimal minimal intervention sets in chapter 4 does not consider graphs with non-manipulative variables, which are indeed prevalent in real-world settings. Notably, Lee and Bareinboim, 2019 explore this direction in the single-objective case, and adapting their approach to the multi-objective setting would be a valuable contribution for MO-CBO.

Automating MO-CBO. Currently, our approach for solving MO-CBO problems requires manually identifying possibly-optimal minimal intervention sets via the graphical characterisation from Theorem 4.1, before applying CAUSAL PARETOSELECT. Automating this process into a single pipeline would enable practitioners to utilise our method out-of-the box for arbitrary graphs. To this end, an algorithm that yields all possibly-optimal minimal intervention sets in the multi-objective setting is required. Such an algorithm can be derived by iterating over all minimal intervention sets and selecting those that meet the criteria from Theorem 4.1.

MOBO algorithm. CAUSAL PARETOSELECT relies on a multi-objective Bayesian optimisation algorithm (MOBO) for solving the local MO-CBO problems. In its current implementation, we use DGEMO (Konakovic Lukovic et al., 2020), which has demonstrated effectiveness in exploring diverse regions of a Pareto front. However, this restricts CAUSAL PARETOSELECT to a single MOBO algorithm, limiting its flexibility to problem specific circumstances. Notably, the choice of MOBO algorithm can and should be adapted based on the specific requirements of the given problem. For instance, certain real-world scenarios involve high levels of noise in function evaluations, which require specialised methods designed for noisy optimisation, e.g. qNEHVI (Daulton et al., 2021). This limitation aligns with the broader need for automation in MO-CBO: Developing an automated, modular framework that supports the selection of different MOBO algorithms would make CAUSAL PARETOSELECT more robust and versatile in practical applications.

8 Conclusion

In this thesis, we have introduced a novel problem formulation called *multi-objective causal Bayesian optimisation* (MO-CBO), for optimising multiple target variables within a causal system by sequentially performing interventions. It extends a line of research called *causal Bayesian optimisation* (CBO) (Aglietti et al., 2023) to the multi-objective setting, incorporating elements of causality theory for causally-informed optimisation. Causal Bayesian optimisation is an active research field, and there is a growing body of literature focusing on the advancement of the CBO method. This thesis contributes to the field by formalising a multi-objective CBO variant with a methodology for solving MO-CBO problems. In the broader sense, our work also belongs to the fields of causal decision-making and causal AI as outlined in chapter 3.

Notably, we aim to solve MO-CBO problems by identifying the causal Pareto front which is given by Pareto-optimal solutions in terms of intervention set-value pairs. Applying a traditional multi-objective Bayesian optimisation (MOBO) algorithm to a MO-CBO problem would perform interventions on all treatment variables, neglecting the causal information available. We have demonstrated that this approach can fail to discover the true causal Pareto front leading to suboptimal solutions. Our proposed algorithm, Causal ParetoSelect, overcomes this limitation by incorporating causality: For a given causal graph, first a set of possibly-optimal minimal intervention sets is identified, among which Pareto-optimal solutions with respect to intervention set-value pairs are searched. Causal ParetoSelect demonstrates improved performance as compared to traditional MOBO on a range of performance indicators, such as cost-efficiency, solution diversity, and solution accuracy.

There are various directions for future research on MO-CBO. The adaptation of existing CBO variants to the multi-objective case is a viable contribution. For instance, combining *dynamic* CBO (Aglietti et al., 2021) with MO-CBO would lead to a MO-CBO variant that can handle time-dynamic causal models. Another exploratory direction is the development of a MO-CBO variant with a Gaussian process upper-confidence-bound acquisition function (Srinivas et al., 2010) that can establish theoretical guarantees for the identified solutions. As the fields of causal decision-making and causal AI continue to grow, we anticipate more progress in the development of multi-objective frameworks to address complex, real-world challenges.

List of Figures

List of Tables

Bibliography

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Foerster, J. N., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, *86*(1), 97–106.

Močkus, J. (1975). On bayesian methods for seeking the extremum. *Optimization Techniques IFIP Technical Conference Novosibirsk*, 400–404.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, *104*, 148–175.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1st). Basic Books, Inc.

Buesing, L., Weber, T., Zwols, Y., Heess, N., Racaniere, S., Guez, A., & Lespiau, J.-B. (2019). Woulda, coulda, shoulda: Counterfactually-guided policy search. *International Conference on Learning Representations*.

Berrevoets, J., Kacprzyk, K., Qian, Z., & van der Schaar, M. (2023). Causal deep learning. *arXiv preprint arXiv:2303.02186*.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*.

Zeitler, J., & Astudillo, R. (2024). Causal elicitation for bayesian optimization. *9th Causal Inference Workshop at UAI 2024*.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, *15*(5).

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal machine learning: A survey and open problems.

Ren, S., & Qian, X. (2024). Causal bayesian optimization via exogenous distribution learning.

Cochran, W., & Cox, G. (1957). *Experimental designs*. Wiley.

Lu, C., Schölkopf, B., & Hernández-Lobato, J. M. (2018). Deconfounding reinforcement learning in observational settings.

Konakovic Lukovic, M., Tian, Y., & Matusik, W. (2020). Diversity-guided multi-objective bayesian optimization with batch evaluations. *Advances in Neural Information Processing Systems*, *33*, 17708–17720.

Zitzler, E., & Thiele, L. (1999). Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, *3*(4), 257–271.

Hillermeier, C. (2001). Generalized homotopy approach to multiobjective optimization. *Journal of Optimization Theory and Applications*, *110*(3), 557–583.

Riquelme, N., Von Lücken, C., & Baran, B. (2015). Performance metrics in multi-objective optimization. *2015 Latin American Computing Conference (CLEI)*, 1–11.

Schulz, A., Wang, H., Grinspun, E., Solomon, J., & Matusik, W. (2018). Interactive exploration of design trade-offs. *ACM Transactions on Graphics*, *37*(4).

Huang, Y., & Valtorta, M. (2006). Pearl's calculus of intervention is complete. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 217–224.

Shpitser, I., & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, 1219–1226.

Tian, J., & Pearl, J. (2002). On the testable implications of causal models with hidden variables. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, 519–527.

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*, 455–492.

Couckuyt, I., Deschrijver, D., & Dhaene, T. (2014). Fast calculation of multiobjective probability of improvement and expected improvement criteria for pareto optimization. *Journal of Global Optimization*, *60*, 575–594.

Ferro, A., Pina, F., Severo, M., Dias, P., Botelho, F., & Lunet, N. (2015). Use of statins and serum levels of prostate specific antigen. *Acta Urológica Portuguesa*, *32*.

Schutze, O., Esquivel, X., Lara, A., & Coello, C. A. C. (2012). Using the averaged hausdorff distance as a performance measure in evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, *16*(4), 504–522.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010a). Gaussian process optimization in the bandit setting: No regret and experimental design. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 1015–1022.

Rasmussen, C. E. (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning: Ml summer schools 2003, canberra, australia, february 2 - 14, 2003, tübingen, germany, august 4 - 16, 2003, revised lectures* (pp. 63–71). Springer Berlin Heidelberg.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning* (pp. 63–71). Springer.

Anonymous. (2024). Causal bayesian optimization with unknown causal graphs [under review]. *Submitted to The Thirteenth International Conference on Learning Representations*.

Imbens, G., & Rubin, D. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Miettinen, K. (1999). *Nonlinear multiobjective optimization* (Vol. 12). Springer Science & Business Media.

Laumanns, M., & Ocenasek, J. (2002). Bayesian optimization algorithms for multi-objective optimization. *Parallel Problem Solving from Nature — PPSN VII*, 298–307.

Zhang, Q., & Li, H. (2007). Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation*, *11*(6), 712–731.

Zhang, Q., Liu, W., Tsang, E., & Virginas, B. (2010). Expensive multiobjective optimization by moea/d with gaussian process model. *IEEE Transactions on Evolutionary Computation*, *14*(3), 456–474.

Gambier, A., & Badreddin, E. (2007). Multi-objective optimal control: An overview. *IEEE International Conference on Control Applications*, 170–175.

Stewart, T., Bandte, O., Braun, H., Chakraborti, N., Ehrgott, M., Göbelt, M., Jin, Y., Nakayama, H., Poles, S., & Di Stefano, D. (2008). Real-world applications of multiobjective optimization. In *Multiobjective optimization: Interactive and evolutionary approaches* (pp. 285–327). Springer Berlin Heidelberg.

Bonilla, E. V., Chai, K., & Williams, C. (2007). Multi-task gaussian process prediction. *Advances in Neural Information Processing Systems*, *20*.

Lee, S., & Bareinboim, E. (2018). Structural causal bandits: Where to intervene? *Advances in Neural Information Processing Systems*, *31*.

Aglietti, V., Dhir, N., González, J., & Damoulas, T. (2021). Dynamic causal bayesian optimization. *Advances in Neural Information Processing Systems*, *34*, 10549–10560.

Xia, K., Lee, K.-Z., Bengio, Y., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, *34*, 10823–10836.

Bareinboim, E., Forney, A., & Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, *28*.

Lattimore, F., Lattimore, T., & Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, *29*.

Deb, K., Agrawal, S., Pratap, A., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, *6*, 182–197.

Zuluaga, M., Sergent, G., Krause, A., & Püschel, M. (2013). Active learning for multi-objective optimization. *Proceedings of the 30th International Conference on Machine Learning*, *28*(1), 462–470.

Knowles, J. D. (2006). Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, *10*, 50–66.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Henrández-Lobato, J. M., Hoffman, M. W., & Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, 918–926.

Aglietti, V., Lu, X., Paleyes, A., & González, J. (2020). Causal bayesian optimization. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, *108*, 3155–3164.

Aglietti, V., Malek, A., Ktena, I., & Chiappa, S. (2023). Constrained causal Bayesian optimization. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 304–321.

Branchini, N., Aglietti, V., Dhir, N., & Damoulas, T. (2023). Causal entropy optimization. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, *206*, 8586–8605.

Gultchin, L., Aglietti, V., Bellot, A., & Chiappa, S. (2023). Functional causal Bayesian optimization. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, *216*, 756–765.

Lee, S., & Bareinboim, E. (2019). Structural causal bandits with non-manipulable variables. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 4164–4172.

Daulton, S., Balandat, M., & Bakshy, E. (2021). Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in Neural Information Processing Systems*, *34*, 2187–2200.

Richens, J., & Everitt, T. (2024). Robust agents learn causal world models. *The Twelfth International Conference on Learning Representations*.

Sussex, S., Sessa, P. G., Makarova, A., & Krause, A. (2024). Adversarial causal bayesian optimization. *International Conference on Learning Representations*.

Sussex, S., Makarova, A., & Krause, A. (2023). Model-based causal bayesian optimization. *The Eleventh International Conference on Learning Representations*.

Peters, J., Janzing, D., & Schlkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.

Bradford, E., Schweidtmann, A. M., & Lapkin, A. (2018). Efficient multiobjective optimization employing gaussian processes, spectral sampling and a genetic algorithm. *Journal of Global Optimization*, *71*(2), 407–438.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010b). Gaussian process optimization in the bandit setting: No regret and experimental design. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 1015–1022.

Belakaria, S., Deshwal, A., Jayakodi, N. K., & Doppa, J. R. (2020). Uncertainty-aware search framework for multi-objective bayesian optimization. *ArXiv*, *abs/2204.05944*.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2). MIT press Cambridge, MA.