Your Assumed DAG is Wrong
And Here's How To Deal With It

Kirtan Padh[*] KIRTAN.PADH@TUM.DE
Zhufeng Li[*] ZHUFENG.LI@HELMHOLTZ-MUNICH.DE
Cecilia Casolo CECILIA.CASOLO@HELMHOLTZ-MUNICH.DE
Niki Kilbertus NIKI.KILBERTUS@TUM.DE
Technical University of Munich, Helmholtz Munich, Munich Center for Machine Learning (MCML)

Editors: Biwei Huang and Mathias Drton

Abstract

Assuming a directed acyclic graph (DAG) that represents prior knowledge of causal relationships between variables is a common starting point for cause-effect estimation. Existing literature typically invokes hypothetical domain expert knowledge or causal discovery algorithms to justify this assumption. In practice, neither may propose a single DAG with high confidence. Domain experts are hesitant to rule out dependencies with certainty or have ongoing disputes about relationships; causal discovery often relies on untestable assumptions itself or only provides an equivalence class of DAGs and is commonly sensitive to hyperparameter and threshold choices. We propose an efficient, gradient-based optimization method that provides bounds for causal queries over a collection of causal graphs—compatible with imperfect prior knowledge—that may still be too large for exhaustive enumeration. Our bounds achieve good coverage and sharpness for causal queries such as average treatment effects in linear and non-linear synthetic settings as well as on real-world data. Our approach aims at providing an easy-to-use and widely applicable rebuttal to the valid critique of 'What if your assumed DAG is wrong?'.

Keywords: causal inference, graphical model, cause-effect estimation, bounding

1. Introduction

Estimating the strength of causal effects is crucial across numerous domains, as it enables informed decision-making and policy interventions. Causal inference techniques have been applied in a wide variety of fields including but not limited to economics (Imbens and Wooldridge, 2009; Athey and Imbens, 2017), education (LaLonde, 1986; Sewell and Shah, 1968; Dehejia and Wahba, 1999), healthcare (Badri et al., 2009; Sanchez et al., 2022), biology (Lecca, 2021), medicine (Castro et al., 2020; Liu et al., 2022; Feuerriegel et al., 2024), and software engineering (Siebert, 2023).

The gold standard for quantifying causal relationships is via active experimentation and intervention, such as in (controlled) clinical trials. However, in many settings, such experiments are unethical or impossible. For instance, randomly assigning smoking habits to assess the effect on lung cancer risk would be highly unethical. Similarly, questions like whether and how strongly the health of institutions affects economic growth cannot feasibly be tackled by active experimentation (Acemoglu et al., 2001). These challenges, combined with the rapid development of computational data analysis

[*] Equal contribution, Code: `https://github.com/zhufengli/your_dag_is_wrong`

methods in the last few decades, have led to significant advancements in techniques to estimate causal effects from observational data (Guo et al., 2020; Malinsky and Danks, 2018; Peters et al., 2017; Glymour et al., 2019).

Within many causal inference frameworks, estimating the strength of causal effects requires knowledge of the underlying causal graph. Specifically, the Structural Causal Model (SCM) (Pearl, 2009a; Peters et al., 2017) framework encodes causal assumptions in the form of Directed Acyclic Graphs (DAGs), where edges represent the data-generating mechanisms in terms of which variables influence others (Pearl, 1995; Lauritzen, 2001). This information, together with observational and/or experimental data, can then be used for downstream cause-effect estimation tasks. However, in practice, the causal graph is often not known with certainty. Most literature on cause-effect estimation assumes that the causal graph is given, typically referring to domain expert knowledge or causal discovery algorithms to justify this assumption. Yet, it is unlikely that even domain experts would assign high confidence to a single graph; they may be more inclined to debate the absence or presence of specific edges rather than agree on a complete structure.

Causal discovery algorithms aim to learn the causal structure from data (Spirtes et al., 2000, 1995; Zheng et al., 2018; Chickering, 2002) (see Vowels et al. (2022) for a more comprehensive review). Despite tremendous efforts, causal discovery has not advanced to a point where it is commonly relied upon in practice. So-called ‘constraint-based’ algorithms rely on conditional independence tests (with high-dimensional conditional sets) (Spirtes et al., 1995, 2000). The difficulty of conditional independence testing (Shah and Peters, 2020; Lundborg et al., 2022) and adaptively compounding errors often render these methods unreliable in practice. Even with a conditional independence oracle, they can only identify causal graphs up to a ‘Markov equivalence’, i.e., their output is typically a collection of DAGs instead of a single one. Other methods for causal discovery critically rely on other untestable assumptions, such as specific function classes or distributional assumptions (Peters et al., 2014). So-called ‘score-based methods’, the third major class of causal discovery algorithms typically require solving non-convex constrained optimization problems, and the proposed DAG critically depends on heuristic hyperparameter choices (Zheng et al., 2018; Wei et al., 2020). Finally, more recent methods acknowledge that committing to a single DAG is overly restrictive and learn entire distributions over plausible graphs (see Mamaghan et al., 2024 for an overview). Ultimately, all these methods are practically limited in that there remains room for uncertainties in the final output (Reisach et al., 2021; Ganian et al., 2024; Sondhi and Shojaie, 2019).

As a result, the assumed causal graph is an elusive object, and there is often ambiguity about various edges. Estimating causal effects based on a single assumed graph is, therefore, likely equally faulty. In this work, we address the uncertainty inherent in the causal graph structure, particularly concerning the existence and direction of certain edges. This uncertainty could arise from underperforming structure learning algorithms or limited domain knowledge. For instance, in large graphs, uncertainty may manifest in the form of cluster DAGs (Anand et al., 2023), where there is knowledge about clusters of nodes and inter-cluster relationships, but no information on intra-cluster causal connections. Alternatively, domain experts might be confident about the existence or non-existence of only certain edges in the graph. We propose an efficient, gradient-based optimization method that computes bounds on cause-effect estimates over all plausible DAGs compatible with the available prior knowledge. Our method is applicable even when the collection of compatible graphs is too large for exhaustive enumeration. Through this approach, we aim to provide a robust framework for causal inference that acknowledges the uncertainty in the assumed causal structure, addressing the critical concern of "What if your assumed DAG is wrong?". Our contributions are as follows:

- We develop a method to efficiently bound causal effects over a (potentially large) flexibly pre-defined set of plausible causal graphs.
- Our method is applicable in non-linear and continuous settings and for large graphs.
- Beyond empirical evaluation on synthetic data, we demonstrate our method on a real-world dataset, where we attempt to find plausible graphs via constraint-based causal discovery.

2. Problem setting and assumptions

2.1. Background and setup

We operate within the fully observed structural causal model (SCM) framework (Pearl, 2009b; Peters et al., 2017), where an SCM is a tuple $(\boldsymbol{V}, \mathcal{F}, N, P_{\boldsymbol{N}})$. Here, $\boldsymbol{V} = \{X_1, \ldots, X_d\}$ are the endogenous variables, $\boldsymbol{N} = (N_1, \ldots, N_d)$ are exogenous (or noise) variables, $P_{\boldsymbol{N}}$ is a distribution over $\boldsymbol{N}$ where the $\boldsymbol{N}$ are jointly independent, and $\mathcal{F} = \{f_i\}_{i=1}^d$ is a collection of functions (the structural equations) such that $X_i := f_i(\mathrm{pa}_{\mathcal{G}}(X_i), N_i)$ for all $i \in \{1, \ldots, d\}$. The set $\mathrm{pa}_{\mathcal{G}}(X_i) \subset \boldsymbol{V}$, called 'the parents of X_i,' is such that the graph $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ over $\boldsymbol{V}$ is acyclic. Here, $\boldsymbol{E} \subset \{(X_i, X_j) \in \boldsymbol{V} \times \boldsymbol{V} \mid i \neq j\}$ is the set of edges. We often denote the DAG induced by the SCM simply by $\mathcal{G}$ when not explicitly referring to the vertices/edges, write $\mathrm{pa}(X_i)$ when the graph is clear from the context, and use $X_i \to X_j$ for $(X_i, X_j) \in \boldsymbol{E}$. Using Pearl's 'do notation' (Pearl, 2009a), $do(X_i = x^*)$ denotes a (hard) intervention on $X_i \in \boldsymbol{V}$ fixing its value to x^*. Technically, this means replacing $f_i \in \mathcal{F}$ with the function $X_i := x^*$. Any fully observed SCM as defined above induces a unique joint distribution $P_{\boldsymbol{V}}$ over $\boldsymbol{V}$, called 'observational distribution,' which is Markov with respect to the graph $\mathcal{G}$, i.e., each $X_i \in \boldsymbol{V}$ is conditionally independent of its non-descendants given its parents (Peters et al., 2017).[1] We will also represent a graph $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ via its adjacency matrix $A_{\mathcal{G}} \in \{0, 1\}^{d \times d}$, where $A_{ij} = 1$ if and only if $(i, j) \in \boldsymbol{E}$. Finally, let $\mathcal{D} = \{x_1^i, x_2^i, \ldots, x_d^i\}_{i=1}^N$ be a dataset of N i.i.d. samples from the observational distribution. In short, we assume data to be generated by a fully observed SCM, but a priori make no assumptions on the distribution $P_{\boldsymbol{N}}$ or the functional relationships $\mathcal{F}$. In particular, we will consider linear and non-linear $\mathcal{F}$ separately later.

Let $\mathcal{E}_{\boldsymbol{V}}$ be the set of all possible edge sets that form a valid DAG with vertices $\boldsymbol{V}$. The set of allowed DAGs can be noted as $\mathcal{G}(\boldsymbol{V}, \mathcal{E})$. The setting we are interested in is when we have partial information about causal relations between some of the variables, i.e., sure edges $\boldsymbol{E}_s$ and forbidden edges $\boldsymbol{E}_f$. $\boldsymbol{E}_s$ are edges whose presence we are certain about, and $\boldsymbol{E}_f$ are edges that we are certain are not present. Let $\mathcal{E}_s \subseteq \mathcal{E}_{\boldsymbol{V}}$ denote the set of edge sets that contain edges in $\boldsymbol{E}_s$. Similarly, let $\mathcal{E}_{\neg f} \subseteq \mathcal{E}_{\boldsymbol{V}}$ denote the set of edge sets that do not contain $\boldsymbol{E}_f$. With this information given, the set of possible DAGs is reduced to $\mathcal{G}(\boldsymbol{V}, \mathcal{E}_s \cap \mathcal{E}_{\neg f})$.

Goal: Our goal is to estimate bounds on a causal query $\mathcal{Q}_{\mathcal{G}}$ based on observational data $\mathcal{D}$, over all possible graphs $\mathcal{G}(\boldsymbol{V}, \mathcal{E}_s \cap \mathcal{E}_{\neg f})$. For instance, common natural queries include interventional expectations of the form $\mathcal{Q}_{\mathcal{G}} = \mathbb{E}[Y \mid do(X = x^*)]$ for $X, Y \in \boldsymbol{V}$. A common way to identify

[1] We assume throughout that densities exist and all random variables have finite variance. Under these mild assumptions, the stated *local* Markov property is equivalent to the observational distribution factorizing according to $\mathcal{G}$ as well as to the global Markov property, namely that d-separations in $\mathcal{G}$ imply conditional independencies in the observational distribution (Peters et al., 2017).

interventional probabilities is via the adjustment formula, where given a valid adjustment set $\boldsymbol{Z} \subset \boldsymbol{V}$, we have (assuming the existence of densities throughout)

$$p\big(y \,|\, do(X = x^*)\big) = \int p(y \,|\, x^*, \boldsymbol{z}) p(\boldsymbol{z}) \, d\boldsymbol{z} \,. \tag{1}$$

This adjustment formula expresses$\mathcal{Q}_G$ only in terms of the observational distribution, i.e., only relying on observed marginal and conditional densities (Pearl, 2009b) as described in Sec. 4.3. In the fully observed setting, every interventional distribution is identified from the observational distribution when the ground truth graph $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ is known exactly. However, when the graph is not known with certainty we typically cannot estimate $\mathcal{Q}_G$ consistently. Yet, we can obtain lower and upper bounds for the quantity over the set of allowed DAGs. There is a trivial brute-force approach, which is to estimate the target query for every possible graph and take the minimum (maximum) over all these values. Since the space of DAGs grows quickly as the number of nodes increases, this quickly becomes computationally infeasible. Instead, we aim to use continuous optimization in a targeted search for the bounds, i.e., the smallest (largest) possible values of the causal query of interest over the space of DAGs. We note that while the underlying quantity is a bound set rather than an interval, we only report the minimum and maximum values from this set.

2.2. Related Work

Point identification of causal effects generally requires strong assumptions on the causal structure or the functional relationships between the variables. When these assumptions are not met, effects can often still be *partially identifiable*, meaning that we can confine the effect to lie within some non-empty, but constrained *set* of effects that are compatible with the structural assumptions and the observed data (Manski, 1990). Such effect sets are often characterized by intervals, defined by a lower and upper bound on the target quantity. Initial work on this was pioneered by Balke and Pearl (1997); Chickering and Pearl (1996), and the topic has attracted much attention lately.

Most of this work has focused on partial identification or sensitivity analysis in the presence of confounding. Different streams of work have focused on bounding causal effects under a variety of assumptions, including assuming discrete domains (Zhang and Bareinboim, 2021; Duarte et al., 2023; Raichev et al., 2024), assuming the presence of an instrument (Gunsilius, 2019; Kilbertus et al., 2020), using neural networks (Hu et al., 2021; Padh et al., 2023), or performing causal sensitivity analysis (Frauen et al., 2024; Melnychuk et al., 2024; Jesson et al., 2021; Marmarelis et al., 2023). Other sources of uncertainty in causal inference include ones arising from interaction patterns in non-IID data (Zhang et al., 2023; Bhattacharya et al., 2020), shifts in covariates (Jesson et al., 2020), and unobserved confounders (Tchetgen Tchetgen, 2013).

Despite these efforts, the uncertainty related to the causal graph itself, particularly in the final goal of treatment effect estimation, has received less attention. Some works focus on improving the inference under the uncertainty introduced by previous causal discoveries on the same data (Chang et al., 2024; Gradu et al., 2024; Malinsky, 2024). For constraint-based causal discovery, these are often characterized by Markov equivalence class (MEC) over which one can perform partial identification (Maathuis et al., 2009; Bellot, 2024). However, the prevailing approach—focusing on uncertain orientations of unoriented edges in the completed partially directed acyclic graph (CPDAG) corresponding to the MEC—is limited for two primary reasons. First, adopting a more flexible approach to managing uncertainty in causal graphs could prove beneficial, especially when

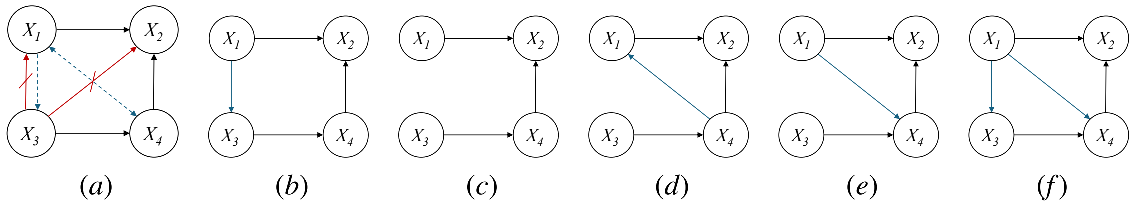


Figure 1: (a) An example illustrating a graph with partial information. Red edges represents forbidden edges and blue dotted edges represent unknown edges. (b)–(f) All plausible DAGs compatible with the information provided in (a).

leveraging domain knowledge to infer edge information. Second, several studies have highlighted the potentially unreliable performance of (constraint-based) causal discovery algorithms in real-world settings (Reisach et al., 2021). Hence, uncertainties beyond the ones arising purely due to causal discovery should be considered. In orthogonal work, Henckel et al. (2024) highlight that the "utility" of a specific assumed graph typically depends on the downstream task the graph is used for. Starting from a downstream task, i.e., cause-effect estimation, they measure how 'close any given DAG is to the ground truth DAG' by comparing how many correct cause-effect estimands the proposed DAG recovers. This idea leads to novel metrics to measure the quality of causal discovery *with a downstream task in mind.*

More recent work by Strieder and Drton (2023, 2024) has developed confidence regions for total causal effects that account for uncertainties in both the causal structure and numerical size of nonzero effects, but their applicability is restricted to linear models. Additionally, efforts to mitigate uncertainties from imperfect expert input have utilized Large Language Models (Long et al., 2023) and pre-processing techniques (Oates et al., 2017).

Our work is complementary to these efforts and aims to further fill the gap in inference under the uncertainty coming from the specification of the causal graph. Unlike previous efforts, our method allows a pre-defined set of plausible causal graphs that are more general than the MEC, does not require the linearity assumption, and uses continuous optimization to allow the method to extend more easily to larger graphs. We describe our method in the next section.

3. General optimization formulation

This section provides a high-level overview of our approach, followed by a more detailed description of the implementation. With the notation introduced in Sec. 2, we can define the unknown edge set as $\boldsymbol{E}_u$ and assume its cardinality to be $K \in \mathbb{N}$. We denote by A_α a $d \times d$ matrix where the entries corresponding to the uncertain edges in the adjacency matrix A are replaced by $\{\alpha_1, \alpha_2, \dots \alpha_K\}$.

3.1. Illustrative example

Fig. 1(*a*) shows an example, where the sure edge set is $\boldsymbol{E}_s = \{(X_1, X_2), (X_3, X_4), (X_4, X_2)\}$ and the forbidden edge set is $\boldsymbol{E}_f = \{(X_3, X_1), (X_3, X_2)\}$. With the above knowledge, we are still unsure about the existence of the edge between X_1 and X_3, as well as between X_1 and X_4, even in terms of their direction. Therefore, we have $\boldsymbol{E}_u = \{(X_1, X_3), (X_1, X_4), (X_4, X_1), \emptyset\}$ and thus $K = 3$. Assuming the quantity $\mathcal{Q}$ we are interested in is $\mathbb{E}[X_2 \mid do(x_3^*)]$, computing $\mathcal{Q}$ depends on the (non)existence of these unknown edges. For the same observed data, we, therefore, have more than one possible value for $\mathcal{Q}$ given the different possible graph structures shown in Figs. 1(b) to 1(f). The minimum and maximum of these values form our bounds. Estimating $\mathcal{Q}$ could involve estimating and combining conditional and marginal densities from observed data in the general case. However, cause effect estimation can often be phrased as a combination rather simple 'function fitting' steps in specific cases, as we will describe in Sec. 3.2.1. More details on further possible causal queries can be found in Appendix A. The following matrix represents the adjacency matrix for the graph shown in Fig. 1(a) including its uncertain entries (we use row/column indices for the αs instead of a single running index for readability)

$$A_\alpha = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 0 & 1 & \alpha_{13} & \alpha_{14} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \alpha_{41} & 1 & 0 & 0 \end{pmatrix} \end{array} . \tag{2}$$

Not all combinations of edges in $\boldsymbol{E}_u$ form a plausible graph because of the acyclicity constraint. In this example, our search space $\mathcal{E}_s \cap \mathcal{E}_{\neg f} = \boldsymbol{E}_s \cup \{\{(X_1, X_3), (X_1, X_4)\}, \{(X_1, X_3)\}, \{(X_1, X_4)\}, \{(X_4, X_1)\}, \{\emptyset\}\}$ is small enough for the brute force approach, but we will show how to get these bounds using continuous optimization. Let $\hat{\mathcal{Q}}$ be the estimated $\mathcal{Q}$ from observed data. We are then looking at the following optimization problem to obtain the lower / upper bounds (for $K = 3$)

$$\min_{\alpha \in \{0,1\}^K} / \max_{\alpha \in \{0,1\}^K} \hat{\mathcal{Q}}(A_\alpha, \mathcal{D}) \,, \qquad \text{subject to } \mathcal{G}(A_\alpha) \in \text{DAGs} \,, \tag{3}$$

where we 'flatten' the variable edges of A_α into a vector $\alpha = (\alpha_{13}, \alpha_{14}, \alpha_{41}) \in \{0, 1\}^K$ in arbitrary order. In words, find the minimum and maximum estimates of $\mathcal{Q}$ over all values of α for which A_α represents a DAG. We note that the search space grows super-exponentially in K, the number of uncertain edges. Due to the combinatorial nature of the problem, global optimization that is more efficient than a brute force search is not straight forward. We tackle this challenge by leveraging similarities to the optimization formulation underlying score-based continuous causal discovery methods. In particular, we show how to build on concepts developed by Zheng et al. (2018) to convert the constrained discrete optimization into an unconstrained continuous optimization problem for which we can use augmented Lagrangian techniques combined with differentiable DAG sampling approaches proposed by, e.g., Charpentier et al. (2022), which have also been updated to incorporate prior knowledge (Rittel and Tschiatschek, 2023).

3.2. Solving the general optimization problem

To operationalize this optimization, we first re-formulate Eq. (3) as a continuous unconstrained optimization problem without introducing any parametric assumptions. Fig. 2 illustrates our overall framework. In this subsection and Sec. 4, we motivate and describe all aspects of this diagram one by one. The first step is to provide an estimator of the causal query $\mathcal{Q}$.

3.2.1. ESTIMATING $\mathcal{Q}$

We resort to cause-effect estimation via valid adjustment sets (Perković et al., 2015, 2018; Shpitser et al., 2010) to control for confounding variables. Concretely, let $X, Y \in \boldsymbol{V}$, where X is the 'treatment' and Y is the 'outcome'. One of the primary challenges in cause-effect estimation comes

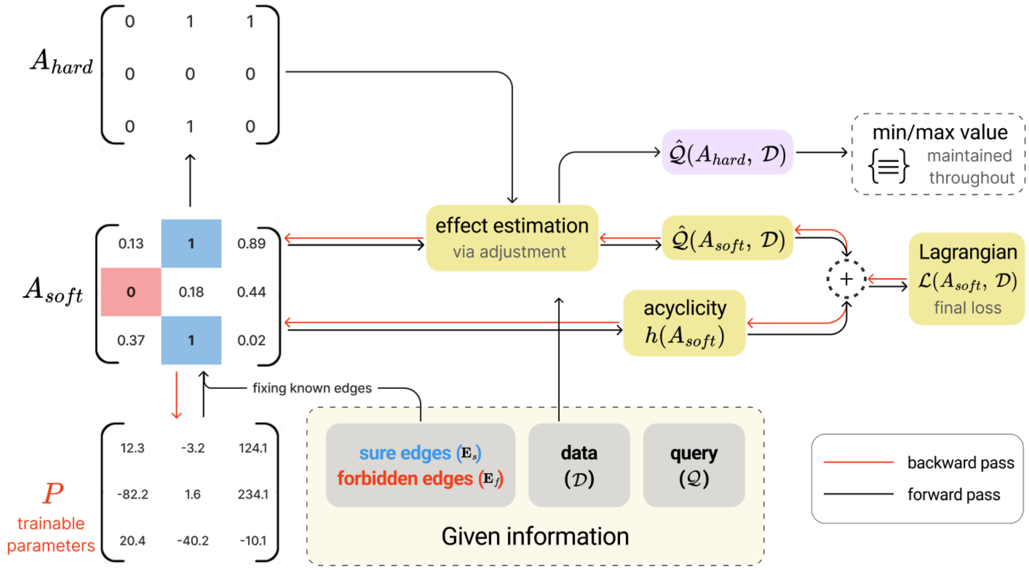


Figure 2: An illustrative diagram of the Lagrangian method.

from confounding variables that influence both the treatment and the outcome. By ‘controlling’ or ‘adjusting’ for these variables, the adjustment set isolates the direct causal effect of the treatment on the outcome from the overall association, which may be due to direct, but also confounding effects. A set $\boldsymbol{Z} \subset \boldsymbol{V}$ is called valid adjustment set for X, Y in the DAG $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ if (i) $\boldsymbol{Z}$ blocks all undirected paths in $\mathcal{G}$ between X and Y, and (ii) $\boldsymbol{Z}$ does not contain any node on a directed path from X to Y in $\mathcal{G}$ or any descendant of such a node (Shpitser et al., 2010). There may be multiple valid adjustment sets for a tuple X, Y and we implement our method using both parent and optimal adjustment sets to estimate the query (Henckel et al., 2022). Selecting these adjustment sets without interfering with the gradient-based continuous optimization involves some non-trivial technicalities. We provide further details on adjustment sets and pseudocode on how we adapt them to fit into our framework in Appendix D.1. Given a valid adjustment set for X, Y in $\mathcal{G}$ allows us to estimate the causal effect of X on Y, i.e., the interventional distribution, via Eq. (1). Naive estimation of this estimand would require conditional density estimation of $p(y \mid x, \boldsymbol{z})$ and an empirical (e.g., sampling-based) integration over the marginal $p(\boldsymbol{z})$. In general, the marginal density could be estimated via a broad range of methods from Gaussian processes, to mixture density network (Bishop, 1994) or conditional normalizing flows/diffusion/flow matching (Papamakarios et al., 2021; Ho et al., 2020; Lipman et al., 2023). Common targeted causal queries such as the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE) can be estimated directly from estimates of the general interventional distribution in Eq. (1). Depending on further parametric assumptions, conditional density estimation can be circumvented and expected causal effect estimation can be reduced more effectively to simple function fitting routines, see Sec. 4.3.

3.2.2. Enforcing acyclicity

Zheng et al. (2018) are widely credited for the introduction of a smooth function of weighted (continuous) adjacency matrices, $h : \mathbb{R}^{d \times d} \to \mathbb{R}_{\geq 0},\ A \mapsto \operatorname{tr}(\exp(A \odot A)) - d$ (where $\odot$ denotes the Hadamard product) that captures the ‘degree of acyclicity’ of the corresponding graph. This function has several desirable properties: (i) It quantifies the degree to which a graph is acyclic referred to as 'DAG-ness' (if $h(A) = 0$, A represents an acyclic graph; the larger $h(A)$ the more 'cyclic' the graph). (ii) It is smooth. (iii) Both the function and its derivative are computationally tractable. By replacing the combinatorial acyclicity constraint ($\mathcal{G}(A_\alpha \in \text{DAGs})$ with the equality constraint $h(A_\alpha) = 0$, the DAG learning problem becomes amenable to conventional continuous constrained optimization techniques based on gradient descent. This enables us to leverage the robust and efficient optimization tools provided by modern machine learning frameworks. Since the original proposal of h by Zheng et al. (2018), various other constraints have been proposed that offer computational or numerical benefits. Vowels et al. (2022, Sec. 5) provide an incomplete list. In this work, we use the constraint defined by Bello et al. (2022), which performs well empirically.

3.2.3. PUTTING IT ALL TOGETHER

With the adjustment based estimand and a continuous equality constraint for acyclicity, what remains is to convert the discrete search space into a continuous one, in which the the query is continuous (and differentiable). Specifically, we would like to replace the search space $\{0, 1\}^K$ in Eq. (3) with $\mathbb{R}^K$. A direct relaxation does not work, as we require strict zeros to indicate the missingness of edges, which in turn affects which $\boldsymbol{Z}$ form valid adjustment sets for a given tuple X, Y. This challenge can be addressed via the Gumble-Softmax straight-through estimator (Jang et al., 2017), which allows us to include discrete distributions as part of an overall differentiable optimization procedure. We describe the details in Sec. 4.1. Finally, putting all these building blocks together allows for a re-formulation of the resulting constrained continuous optimization problem into an unconstrained one using the augmented Lagrangian formulation (Birgin and Martínez, 2014) as detailed in Sec. 4.2.

We should point out that this is far from the only way to tackle this optimization. Any of the array of score-based continuous causal discovery methods (Wei et al., 2020; Montagna et al., 2023; Bello et al., 2022; Charpentier et al., 2022; Zheng et al., 2018) could be used as well if adapted correctly, and we hope that this work serves as a starting point to do so. In particular, in the next sections, we focus on Charpentier et al. (2022); Zheng et al. (2018).

4. Operationalizing the Optimization

4.1. Discrete distributions: Straight-Through Gumbel-Softmax (ST-GS) estimator

Representing a directed graph numerically via a binary adjacency matrix is a natural choice. However, any function of this binary matrix will be non-differentiable and thus forestall the use of backpropagation. To allow gradients to flow through the binary adjacency matrix, required to optimize over graphs, we use the Gumbel-Softmax distribution relaxation (Jang et al., 2017), which approximates categorical samples from continuous variables via a simple Softmax calculation. At a high level, the idea is to bridge the gap between a continuous parameter set and the discrete search space through the use of the Gumble-Softmax distribution. It is called a Gumbel-Softmax straight-through estimator because the forward pass uses a discrete, binary sampled matrix, while the backward pass employs the softmax probabilities derived from the same logits, ensuring differentiability despite the discretization in the forward pass. Following Rittel and Tschiatschek (2023), in practice we optimize over a matrix $P \in \mathbb{R}^{d \times d}$ (see Fig. 2), in which some entries get frozen (sure/forbidden edges) to obtain a 'soft' adjacency matrix A_{soft}. The soft adjacency matrix parameterizes a Gumble-Softmax distribution from which a 'hard' binary adjacency matrix A_{hard} is sampled to represent the actual causal graph

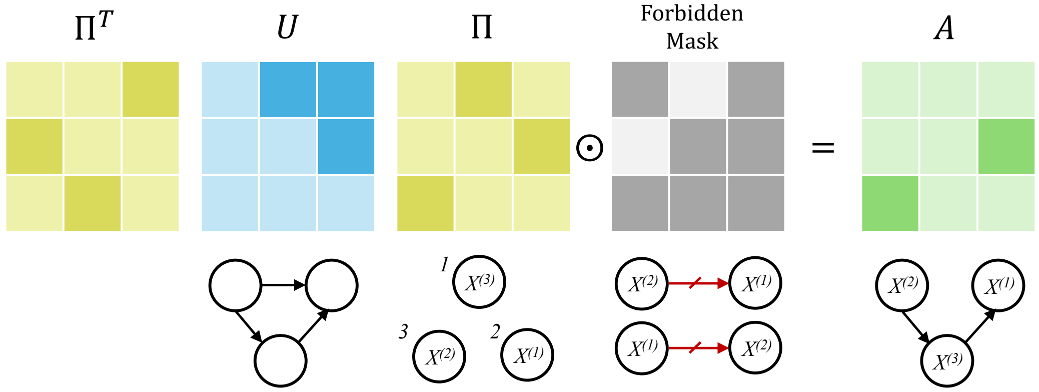


Figure 3: An illustration of the DP-DAG sampling procedure with forbidden edge masking, where dark and light colors represent 1s and 0s, respectively.

$\mathcal{G}$ used in the cause-effect estimation of the forward pass, e.g., to find valid adjustment sets. In the backward pass, the gradients flow through the 'soft' adjacency matrix to the continuous parameters actually being optimized. Details are described in Appendix D.2. For the actual optimization over DAGs we focus on the (augmented) Lagrangian (Zheng et al., 2018) and the DP-DAG (Charpentier et al., 2022) methods.

4.2. Optimizing over the space of DAGs

Lagrangian optimization. Denoting our estimator of the causal query $\mathcal{Q}$ by $\hat{\mathcal{Q}}$ and employing the acyclicity constraint as described in Sec. 3.2, the optimization problem in Eq. (3) becomes

$$\min_{\alpha \in \mathbb{R}^K} / \max_{\alpha \in \mathbb{R}^K} \; \hat{\mathcal{Q}}(A_\alpha, \mathcal{D}) \,, \qquad \text{subject to } h(A_\alpha) = 0 \,. \tag{4}$$

We use the augmented Lagrangian method for (in)equality constraints to solve the constrained optimization problem in Eq. (4) by converting it into an unconstrained problem. The formulation is taken from Nocedal and Wright (2006, Sec. 17.3). Further details are provided in Appendix D.3, which includes the pseudocode for implementing the Lagrangian method in both linear and non-linear settings, presented respectively in Alg. 1 and Alg. 2.

Differentiable DAG sampling. In the augmented Lagrangian method, the constraints are enforced essentially via penalty terms that can rarely be ensured to be exactly zero. Hence, they still sometimes yield cyclic graphs after convergence. Therefore, we additionally apply the DP-DAG method (Charpentier et al., 2022), which provides another solution for sampling DAGs from a trainable distribution. Permitting node indices permutations, any full DAG can be represented as an upper triangular adjacency matrix with zero diagonal and all non-zero entries being 1.

DP-DAG samples a DAG by generating an upper triangular matrix U, which serves as the graph topology, i.e., a plain graph structure without attribution of node indices. In order to attribute indices to the graph, it generates a permutation matrix Π. The permutation is applied on the row and column indices to preserve the graph skeleton, resulting in $A = \Pi^T U \Pi$. We illustrate this process in Fig. 3. Removing edges in the obtained graph amounts to replacing 1 entries with 0s via masking. Removing edges can not introduce directed cycles. In our experiments, we only enforce forbidden edges when using the DP-DAG approach, leading to valid—albeit potentially looser bounds.

4.3. Query: Average effects

While any query $\mathcal{Q}$, i.e., any functional of interventional distributions, can generally be estimated from observational data as described in Sec. 3.2.1, we can significantly simplify the formulation when restricting our attention to the expected value $\mathcal{Q} = \mathbb{E}[Y \mid do(X = x^*)]$ as our query of interest, from which we can directly compute, for example, average treatment effects. Let $\boldsymbol{Z}$ be a valid adjustment set for X, Y in a given DAG $\mathcal{G}$. Then we have

$$\mathcal{Q} = \int y\, p(y \mid do(X = x^*))\, dy = \int y \int p(y \mid x^*, \boldsymbol{z}) p(\boldsymbol{z})\, d\boldsymbol{z}\, dy = \int p(\boldsymbol{z}) \mathbb{E}[Y \mid x^*, \boldsymbol{z}]\, d\boldsymbol{z} \;, \tag{5}$$

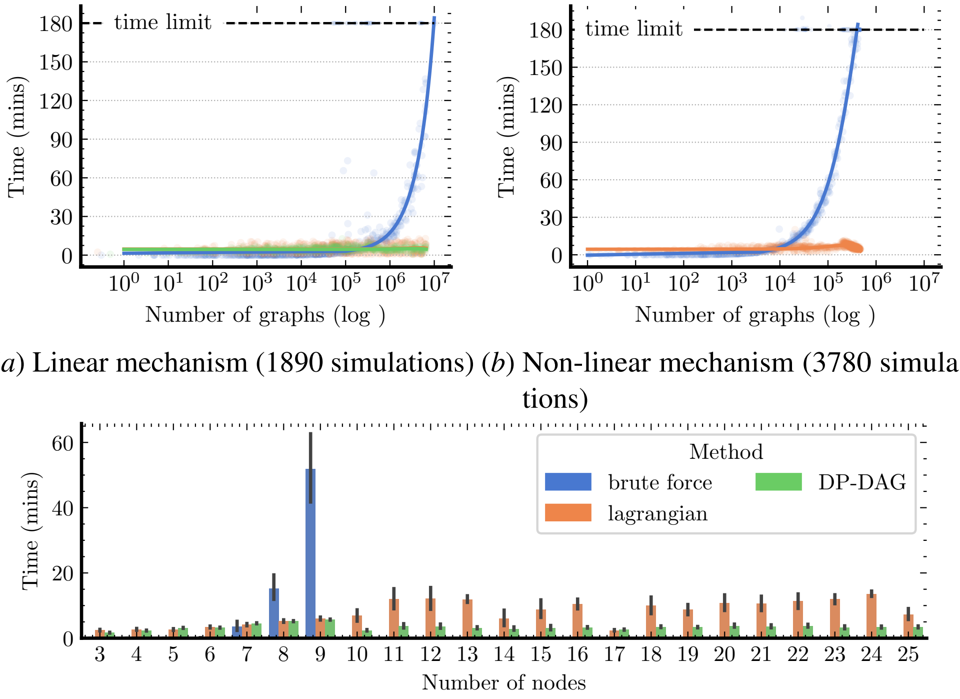
which we can estimate directly from observational data via $\hat{\mathcal{Q}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbb{E}}[Y \mid x^*, \boldsymbol{z}_i]$ for any estimator of the conditional expectation, i.e., any machine learning regressor predicting Y from x and $\boldsymbol{z}$. Depending on the dimensionality and modality of $\boldsymbol{Z}$, we may choose to estimate the density of $\boldsymbol{Z}$, in which case a Monte Carlo approximation of the integral over $\boldsymbol{Z}$ can be computed with sampling access to $\boldsymbol{Z}$. We highlight that the nodes in $\boldsymbol{Z}$, the valid adjustment set, depends on the optimization parameters α. Hence, the dependence of the objective $\mathcal{Q}$ on α, representing the graph, comes solely from what constitutes a valid adjustment set for X, Y. In practice, to evaluate $\mathcal{Q}(A_\alpha, \mathcal{D})$, the query for the current value of α and given data $\mathcal{D}$, we thus only need to fit a single regression function. Hence, our approach is directly applicable to nonlinear contexts, where we use small Multilayer Perceptron (MLP) to estimate $\mathbb{E}[Y \mid x, \boldsymbol{z}]$. In linear settings, we can even sidestep the empirical mean over $\boldsymbol{z}$, by directly reading off the causal effect from the coefficient of x in a single ordinary least squares linear regression from $X, \boldsymbol{Z}$ to Y (Henckel et al., 2022).

5. Experimental results

In this section, we evaluate our algorithm across different synthetic data scenarios, which differ in their data-generating processes and levels of uncertainty. Additionally, we apply our algorithm to the real-world Infant Health and Development Program (IHDP) dataset (Hill, 2011). Hyperparameter choices and specifics of the data-generating processes are in Appendix B. Additional implementation details and information on the packages and data resources used in the implementation can be found in Appendix D and Appendix E respectively.

Data and uncertainty generation. We sample random Erdős–Rényi graphs and use one of the three causal mechanisms (linear, sigmoid) described in Table 3 to generate data based on each graph. Uncertain edges are either obtained (i) *randomly*, by assigning a pre-set proportion of the true edges and non-edges as known uniformly at random, or (ii) *using the PC-algorithm* for causal discovery (Spirtes et al., 2000), running it on different permutations of the variables $\boldsymbol{V}$ and taking the intersection of directed edges and missing edges across the runs as known and all others as unknown. The latter option illustrates challenges in the 'discovery then estimation' approach to causal inference purely from observational data without any additional assumptions.

Comparison metrics. Estimated bounds should contain the ground truth effect while being as informative, i.e., narrow, as possible. Moreover, runtime and scalability is an important aspect in this problem as the reference, brute force search, becomes intractable quickly. This motivates the



(*a*) Linear mechanism (1890 simulations) (*b*) Non-linear mechanism (3780 simulations)

(*c*) Linear mechanism (2322 simulations)

Figure 4: **Runtime comparison**. Our proposed algorithms scale almost constant in the number of graphs in the search space, compared to a super-exponential growth in runtime for the brute force algorithm. (see Table 4 for the details of the data simulations)

following metrics computed individually for each run, denoting true and estimated lower/upper bounds on $\mathcal{Q}$ by l/u and $\hat{l}, \hat{u}$, respectively (while assuming $l \leq u, \hat{l} \leq \hat{u}$).[2]

- **Runtime** is the time elapsed for each method on the same hardware.
- **Point coverage** is a binary indicator whether the true causal effect is covered by the estimated bounds, i.e., $1[\hat{l} \leq \text{ground truth } \mathcal{Q} \leq \hat{u}]$.
- **Bound coverage** measures the proportion of ground truth bounds that are covered by the estimated bounds and is given by $\frac{|[l,u]\cap[\hat{l},\hat{u}]|}{|[l,u]|}$, where $|\cdot|$ denotes the length of intervals (Lebesgue measure on $\mathbb{R}$). This metric lies in $[0, 1]$ with 0 meaning no overlap and 1 meaning that the ground truth bounds are fully contained within the estimated bounds.
- **Bound narrowness** measures how much wider the estimated bounds are compared to the ground truth bounds, which we compute as $\frac{|[\hat{u}-\hat{l}]|}{|[l,u]\cap[\hat{l},\hat{u}]|}$. This metric lies in $[1, \infty)$ where 1 indicates that the estimated bounds are at least as narrow as the ground truth bounds and larger values mean wider estimated bounds. When both bound coverage and narrowness are 1, we perfectly recover the true bounds.

Runtime comparison. Fig. 4 demonstrates that our proposed methods can handle larger, higher-dimensional efficiently with near constant runtime over a wide range of search space sizes. We

[2] In synthetic settings we can obtain ground truth bounds for small graphs by a brute force search over all DAGs compatible with the prior knowledge of forbidden and sure edges.

Mechanism	**Point Coverage**		**Bound Coverage**		**Bound Narrowness**	
(n=5670)	Lagrangian	DP-DAG	Lagrangian	DP-DAG	Lagrangian	DP-DAG
linear	0.97 ± 0.0	$\mathbf{1 \pm 0.0}$	0.93 ± 0.0	$\mathbf{0.99 \pm 0.0}$	$\mathbf{2.46 \pm 1.35}$	2.51 ± 0.12
non-linear	0.99 ± 0.0	N/A	0.93 ± 0.0	N/A	2.24 ± 0.03	N/A

Table 1: Mean values of our key metrics with standard error across all synthetic data simulations where the uncertain edges are chosen randomly.

hypothesize that this near constant scaling comes from the fact that modern gradient-based optimization frameworks (as implemented in PyTorch, jax, etc) are heavily tuned such that an update of $\mathcal{O}(d^2)$ parameters takes roughly equally long for a wide range of d. The brute force approach shows the expected super-exponential increase in computation time with more nodes. Further experiments on runtime comparison are presented in Appendix C.

Bound coverage/narrowness. Table 1 shows point coverage and bound coverage, which are both close to 1 for our approaches. (They are 1 by definition for the brute force search.) The bound narrowness reflects the trade off we have to make for the gained efficiency. Since first-order methods (especially over discrete search spaces) may converge to local optima, our bounds are often wider than they need to be—albeit still valid. Our accounting for finite sample estimation uncertainty in the obtained bounds as discussed in Appendix B.4 plays a role in bound narrowness, as we deliberately and conservatively widen the bounds to account for estimation uncertainty. In cases where the true bounds are narrow, this accounting for finite sample uncertainty can lead to outliers in bound narrowness, substantially influencing the mean values reported here. Bound coverage results for different adjustment set choices and uncertainty sources can be found in Appendix C.

5.1. Real-world experiment: IHDP dataset

We now show how one can approach cause-effect estimation in a real-world setting entirely from observational data without trusting any prior information about the causal structure.

A common approach for cause-effect estimation is to run causal discovery as a first stage and use the resulting graph as input for the cause-effect estimation part. When not willing to make additional assumptions about the functional form of the causal relationships, constraint-based causal discovery algorithms such as the PC algorithm Spirtes et al. (2000) are a common choice. However, the PC algorithm can only ever recover an equivalence class, i.e., we are left with a collection of graphs, not a unique one. Moreover, it has various limitations: the PC algorithm does not scale well to high dimensions Glymour et al. (2019), the required conditional independence tests are hard (Shah and Peters, 2020; Lundborg et al., 2022), error propagation is difficult to handle throughout the algorithm (Harris and Drton, 2013), and the outcome of the PC algorithm (in its standard form) may even depend on the order in which variables are provided in the input (Wienöbst and Liśkiewicz, 2021). Overall, without strong assumptions, it is challenging to obtain a unique causal graph that can be trusted with high confidence from causal discovery methods. Instead, we leverage these inherent uncertainties of the PC algorithm output. We apply the PC algorithm to different permutations of the variables in the dataset, obtaining a Markov equivalence class of graphs each time. We then interpret edges present in all graphs as sure edges, and the ones always missing as forbidden edges in our framework.

Lower Bound	Upper Bound	Ground Truth Effect
2.4 ± 2.01	7.2 ± 2.1	4.7 ± 0.13

Table 2: The estimated bounds and ground truth effect of the IHDP dataset. The errors are the standard deviations to 10 bootstrap sampling iterations with replacement.

We test this approach on the Infant Health and Development Program (IHDP) dataset (Hill, 2011), data from a randomized trial that examined the effects of child care home visits by specialists on low-birth-weight children's future cognitive test scores. Selection bias is introduced to the data artificially by removing a subset of the patients, specifically all children with nonwhite mothers, as in Hill (2011); Shalit et al. (2017). We follow the data prepossessing of Louizos et al. (2017), using the simulated outcomes implemented in the NPCI package (Dorie, 2016) and average over 1000 realizations of the outcomes. The analyzed dataset includes 25 covariates about mothers and children among 747 data samples, among which are 139 treated and 608 control. The study collected numerous pre-treatment variables about the children (such as birth information and neonatal health index) and behavior during pregnancy as well as demographic information on the mothers. For further details on the dataset, see Hill (2011). For causal discovery, we use the PC implementation in `causallearn` package (Zheng et al., 2024) with the kernel-based conditional independence test from Zhang et al. (2011). We applied the PC algorithm to 10 covariates perturbations. We apply our nonlinear bounding framework to this dataset and assess the coverage of the ground truth causal effect, as shown in Table 2.

6. Discussion

In this paper, we address an essential and common critique of DAG-based causal methods: "What if the assumed causal graph is incorrect?" Our method supports more robust conclusions even when the exact causal structure remains uncertain by offering a more computationally efficient way to compute bounds on causal effects across a set of plausible graphs. Moreover, we provided a practical recipe for integrating our approach with causal discovery algorithms, allowing practitioners to manage uncertainties in a methodical alternative to relying on a single estimated graph, adding a layer of reliability to causal analysis. Our experimental results highlight that the method delivers meaningful bounds on causal effects across various graph configurations, including both linear and non-linear settings, and simulated and real-world cases, in a computationally tractable manner.

Limitations and future work: Our method is restricted to fully observed settings with no hidden confounding, limiting its applicability in realistic scenarios where confounding is common. It also relies on a non-convex relaxation of a discrete search space without guaranteed optimal bounds, and the Lagrangian approach does not strictly confine the search to DAGs, though we overcome this through a filter in the optimization process. Moreover, we assume a DAG structure, a premise challenged in literature. Future work could incorporate hidden variables to address unobserved confounding, expand to more general graph structures, adapt the method for approaches beyond Lagrangian and DP-DAG (requiring modifications for sure and forbidden edges), and explore defining and optimizing over alternative representations of uncertainty.

Acknowledgments Cecilia Casolo is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401, 2001.

Tara V Anand, Adele H Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12172–12179, 2023.

Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, 31(2):3–32, 2017.

Masood A Badri, Samaa Attia, and Abdulla M Ustadi. Healthcare quality and moderators of patient satisfaction: testing for causality. *International journal of health care quality assurance*, 22(4): 382–410, 2009.

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

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Alexis Bellot. Towards bounding causal effects under markov equivalence. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244, pages 308–332. PMLR, 2024.

Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pages 1028–1038. PMLR, 2020.

Ernesto G Birgin and José Mario Martínez. *Practical augmented Lagrangian methods for constrained optimization*. SIAM, 2014.

Christopher M Bishop. Mixture density networks. 1994.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, pages 21865–21877. Curran Associates, Inc., 2020.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Ting-Hsuan Chang, Zijian Guo, and Daniel Malinsky. Post-selection inference for causal effects after causal discovery. *arXiv preprint arXiv:2405.06763*, 2024.

Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable dag sampling. In *International Conference on Learning Representations (ICLR)*, 2022.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

David Maxwell Chickering and Judea Pearl. A clinician's tool for analyzing non-compliance. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, pages 1269–1276, 1996.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053–1062, 1999.

Vincent Dorie. Npci: Non-parametrics for causal inference. *URL: https://github. com/vdorie/npci*, 11:23, 2016.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, pages 1–16, 2023.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. *Advances in Neural Information Processing Systems*, 36, 2024.

Robert Ganian, Viktoriia Korchemna, and Stefan Szeider. Revisiting causal discovery from a complexity-theoretic perspective. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 3377–3385. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/374. URL https://doi.org/10.24963/ijcai.2024/374.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Paula Gradu, Tijana Zrnic, Yixin Wang, and Michael I. Jordan. Valid inference after causal discovery. *Journal of the American Statistical Association*, 0(0):1–12, 2024. doi: 10.1080/01621459.2024.2402089. URL https://doi.org/10.1080/01621459.2024.2402089.

Emil Julius Gumbel. Les valeurs extrêmes des distributions statistiques. In *Annales de l'institut Henri Poincaré*, volume 5, pages 115–158, 1935.

Emil Julius Gumbel. The return period of flood flows. *The annals of mathematical statistics*, 12: 163–190, 1941.

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

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with NumPy. *Nature*, 2020.

Naftali Harris and Mathias Drton. Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11), 2013.

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, March 2022. ISSN 1467-9868. doi: 10.1111/rssb.12451. URL http://dx.doi.org/10.1111/rssb.12451.

Leonard Henckel, Theo Würtzen, and Sebastian Weichwald. Adjustment identification distance: a gadjid for causal structure learning. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org, 2024.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12104–12112, 2021.

John D Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 2007.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11637–11649. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/860b37e28ec7ba614f00f9246949561d-Paper.pdf.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pages 4829–4838. PMLR, 2021.

Niki Kilbertus, Matt J Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems*, 33:20108–20119, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv e-prints*, pages arXiv–1412, 2014.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Steffen L Lauritzen. Causal inference from graphical models. *Monographs on Statistics and Applied Probability*, 87:63–108, 2001.

Paola Lecca. Machine learning for causal inference in biological networks: perspectives of this challenge. *Frontiers in Bioinformatics*, 1:746712, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xiaoxuan Liu, Ben Glocker, Melissa M McCradden, Marzyeh Ghassemi, Alastair K Denniston, and Lauren Oakden-Rayner. The medical algorithmic audit. *The Lancet Digital Health*, 4(5): e384–e397, 2022.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*, 2023.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Anton Rask Lundborg, Rajen D Shah, and Jonas Peters. Conditional independence testing in hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850, 2022.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133 – 3164, 2009. doi: 10.1214/09-AOS685. URL https://doi.org/10.1214/09-AOS685.

Daniel Malinsky. A cautious approach to constraint-based causal model selection. *arXiv preprint arXiv:2404.18232*, 2024.

Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

Amir Mohammad Karimi Mamaghan, Panagiotis Tigas, Karla Henrik Johansson, Yarin Gal, Yashas Annadani, and Stefan Bauer. Challenges and considerations in the evaluation of bayesian causal discovery. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

Myrl G Marmarelis, Elizabeth Haddad, Andrew Jesson, Neda Jahanshad, Aram Galstyan, and Greg Ver Steeg. Partial identification of dose responses with hidden confounders. In *Uncertainty in Artificial Intelligence*, pages 1368–1379. PMLR, 2023.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Partial counterfactual identification of continuous outcomes with a curvature sensitivity model. *Advances in Neural Information Processing Systems*, 36, 2024.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *Conference on Causal Learning and Reasoning*, pages 752–771. PMLR, 2023.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Chris J Oates, Jessica Kasza, Julie A Simpson, and Andrew B Forbes. Repair of partly misspecified causal diagrams. *Epidemiology*, 28(4):548–552, 2017.

Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, and Niki Kilbertus. Stochastic causal programming for bounding treatment effects. In *Conference on Causal Learning and Reasoning*, pages 142–176. PMLR, 2023.

The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

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

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009a. doi: 10.1214/09-SS057. URL https://doi.org/10.1214/09-SS057.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jak Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 2011.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. A complete generalized adjustment criterion. In *Uncertainty in Artificial Intelligence-Proceedings of the Thirty-First Conference (2015)*, pages 682–691. AUAI Press, 2015.

Emilija Perković, Johannes Textor, and Markus Kalisch. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18:1–62, 2018.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL http://jmlr.org/papers/v15/peters14a.html.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Anna Raichev, Jin Tian, Alexander Ihler, and Rina Dechter. Estimating causal effects from learned causal networks. In *ECAI 2024*, pages 2524–2531. IOS Press, 2024.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Simon Rittel and Sebastian Tschiatschek. Specifying prior beliefs over dags in deep bayesian causal structure learning. In *ECAI 2023: 26th European Conference on Artificial Intelligence, including 12th Conference on Prestigious Applications of Intelligent Systems, PAIS 2023-Proceedings*, pages 1962–1969. IOS Press, 2023.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

William H Sewell and Vimal P Shah. Social class, parental encouragement, and educational aspirations. *American journal of Sociology*, 73(5):559–572, 1968.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), June 2020. ISSN 0090-5364. doi: 10.1214/19-aos1857. URL http://dx.doi.org/10.1214/19-AOS1857.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

Ilya Shpitser, Tyler VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 527–536, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

Julien Siebert. Applications of statistical causal inference in software engineering. *Information and Software Technology*, page 107198, 2023.

E Smucler, F Sapienza, and A Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, 109(1):49–65, 03 2021. ISSN 1464-3510. doi: 10.1093/biomet/asab018. URL https://doi.org/10.1093/biomet/asab018.

Ezequiel Smucler and Andrea Rotnitzky. A note on efficient minimum cost adjustment sets in causal graphical models, 2022. URL https://arxiv.org/abs/2201.02037.

Arjun Sondhi and Ali Shojaie. The reduced pc-algorithm: improved causal structure learning in large random networks. *Journal of Machine Learning Research*, 20(164):1–31, 2019.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506, 1995.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

David Strieder and Mathias Drton. Confidence in causal inference under structure uncertainty in linear causal models with equal variances. *Journal of Causal Inference*, 11(1):20230030, 2023.

David Strieder and Mathias Drton. Dual likelihood for causal inference under structure uncertainty. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 1–17. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/strieder24a.html.

Eric Tchetgen Tchetgen. The control outcome calibration approach for causal inference with unobserved confounding. *American Journal of Epidemiology*, 179(5):633–640, 12 2013. ISSN 0002-9262. doi: 10.1093/aje/kwt303. URL https://doi.org/10.1093/aje/kwt303.

Guido van Rossum and Fred L Drake. *Python 3 Reference Manual*. CreateSpace, 2009.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Dennis Wei, Tian Gao, and Yue Yu. Dags with no fears: A closer look at continuous optimization for learning bayesian networks. *Advances in Neural Information Processing Systems*, 33:3895–3906, 2020.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Marcel Wienöbst and Maciej Liśkiewicz. An approach to reduce the number of conditional independence tests in the pc algorithm. In *KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44*, pages 276–288. Springer, 2021.

Andy B Yoo, Morris A Jette, and Mark Grondona. Slurm: Simple linux utility for resource management. In *Workshop on job scheduling strategies for parallel processing*, pages 44–60. Springer, 2003.

Chi Zhang, Karthika Mohan, and Judea Pearl. Causal inference under interference and model uncertainty. In *Conference on Causal Learning and Reasoning*, pages 371–385. PMLR, 2023.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12207–12215, 2021.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813. AUAI Press, 2011. ISBN 9780974903972.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

Appendix A. Possible causal queries

While we focus on Average Treatment Effect (ATE), different causal queries could be bounded using this infrastructure. As seen before, the do-notation is employed to describe specific interventional scenarios within a causal model. For instance, the query $p(y \mid do(X = x))$ represents the distribution of the outcome Y following an intervention where X is set to a specific value x, isolating the direct causal effect of X on Y. To assess the impact of varying levels of intervention, comparative queries such as $p(y \mid do(X = x_1))$ versus $p(y \mid do(X = x_2))$ are used. Although we focus on single-variable causal effect estimation, a straightforward application of our framework is when the cause and effect are multiple variables, i.e., $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_{d_Y}\}$ and $\boldsymbol{X} = \{X_1, X_2, \ldots, X_{d_X}\}$, then the query takes the form of $\mathcal{Q} = p(\boldsymbol{y} \mid do(\boldsymbol{X} = \boldsymbol{x}))$. Additionally, to allow for heterogeneity, queries like $p(Y \mid do(X = x), \boldsymbol{Z} = \boldsymbol{z})$ condition on other variables $\boldsymbol{Z}$ (while we use the same symbol, this does not refer to a valid adjustment set here) to look at causal effects within different subgroups defined by $\boldsymbol{Z} = \boldsymbol{z}$. The calculation of differences in expected outcomes, $E[Y \mid do(X = x_1)] - E[Y \mid do(X = x_2)]$, quantifies the effect size of an intervention, typically denoted by the Average Treatment Effect (ATE). Lastly, longitudinal studies might use $p(y \mid do(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k))$ to understand the effects of sequential interventions over time.

Appendix B. Experiment setup

B.1. Data generation

We follow Brouillard et al. (2020) in defining the data-generating process in the following steps: 1. We randomly order the nodes to establish a topological ordering. 2. Edges are chosen according to Erdős–Rényi model while ensuring the causal ordering from step 1. Each potential edge is included in the graph independently with probability p, which controls the density of the graph. 3. For each node, the functional form of its causal mechanism is chosen to be among i)linear additive model ii)sigmoidal mixed model, or iii)sigmoidal additive model. The corresponding parameters are then sampled uniformly from preset ranges. 4. Observations are generated in topological order (from root nodes to leaf nodes). Root node values are drawn from a uniform initial distribution, with Gaussian noise added at a coefficient of 0.1. Each generated observational dataset consists of $5,000$ samples. We summarize the causal mechanisms we are using along with the preset parameter ranges in table Table 3.

B.2. Defining the uncertainty

Once a ground truth graph and the associated data are generated as described above, we get sure and forbidden edges in two different ways.

- **Random choice:** We set a `sure edge probability` that represents the proportion of the edges present in the ground truth graph to be included in the 'sure edges' set and a `forbidden edge probability` that represents the proportion of the edges absent from the ground truth graph to be included in the 'forbidden edges' set.
- **PC algorithm:** We run the PC algorithm with Fisher's test for a number of permutations on the order of the input variables. Each permutation does not usually give the same result as the order in which conditional independence tests are performed in the PC algorithm matters for the final outcome. We take the set of sure edges to be edges that exist in each of the graphs and the set of

Table 3: Data-generating models that we have used in the paper with all parameter ranges

Model Name	**Formulation**	**Parameters Range**
linear additive model	$y = \sum_{i \in \text{pa}} \beta_i x_i + \lambda\epsilon$	$\beta \sim \text{Uniform}([-1, -0.25] \cup [0.25, 1])$
sigmoidal mixed model	$y = \frac{\alpha\beta(\sum_{i \in \text{pa}} x_i + \lambda\epsilon)}{1+\lvert\beta(\sum_{i \in \text{pa}} x_i + \lambda\epsilon + \gamma)\rvert}$	$\alpha \sim \text{Exponential}(0.25) + 1$ $\beta \sim \text{Uniform}([-1, -0.25] \cup [0.25, 1])$ $\gamma \sim \text{Uniform}([-2, 2])$
sigmoidal additive model	$y = \sum_{i \in \text{pa}} \frac{\alpha\beta(x_i + \gamma)}{1+\lvert\beta(x_i + \gamma)\rvert} + \lambda\epsilon$	$\alpha \sim \text{Exponential}(0.25) + 1$ $\beta \sim \text{Uniform}([-1, -0.25] \cup [0.25, 1])$ $\gamma \sim \text{Uniform}([-2, 2])$

forbidden edges to be edges that do not exist in any of the graphs. It is crucial to note here that due to the way the sure and forbidden edges are generated in this setting, the set of possible graphs we get from this method might not contain the ground truth graph that generated the data. This is because there might be an edge in the true graph which is always removed by the PC algorithm or vice versa for a non-existing edge

B.3. Hyperparameter choices

For the Lagrangian optimization, we run the optimization for 100 rounds for graphs up to 6 nodes and 200 rounds for graphs with $7, 8$ or 9 nodes, with 30 optimization steps per round. The learning rate is set to 0.3, and the acyclicity constraint used is the "DAGMA" constraint as defined by Bello et al. (2022). The augmented Lagrangian parameters include an initial λ of 2, an initial τ of 0.1, a maximum τ of 4, and a γ of 1.2 (for the notation as defined in Appendix D.3). For the non-linear case, a multilayer perceptron (MLP) must be learned at each step to estimate the value of the causal query from the data once we have the adjustment set. This MLP has 1 hidden layer of dimension 32, and is trained for 1000 epochs with a learning rate of 0.05 and using the Adam optimizer (Kingma and Ba, 2014). For the probabilistic DAG method, we use the same number of rounds as for the Lagrangian, with a learning rate of 0.0001. For both methods, the initial temperature for Gumbel-Softmax sampling is set at 1, decaying at 0.9997 per step.

All of these hyperparameter settings were maintained consistently across nearly 10,000 simulations. While some real-world applications may require tailored adjustments, these largely untuned parameters reliably produced robust results throughout our diverse experimental settings.

Table 4 shows the parameters for our data generation over which be built a Euclidean grid for our experiments, resulting in 5670 simulations for the case with random uncertainty and 3150 simulations for PC uncertainty.

B.4. Estimation uncertainty

It is possible that the ground truth causal query lies on the boundary among all possible values given known edge information. However, even though we know the true causal graph beforehand, the limited data available introduces considerable uncertainty in our statistical estimates. Therefore,

Parameter	Random uncertainty	PC uncertainty
number of nodes	[3, 4, 5, 6, 7, 8, 9]	[3, 4, 5, 6, 7, 8, 9]
sure edge probability	[0.3, 0.5, 0.7]	N/A
forbidden edge probability	[0.3, 0.5, 0.7]	N/A
causal mechanism	['linear', 'sig add', 'sig mix']	['linear', 'sig add', 'sig mix']
random seed	[0, 17, 34, 51, 68]	[0, 17, 34, 51, 68]
edge probability	[0.3, 0.5, 0.7]	[0.3, 0.5, 0.7]
adjustment type	[parent, optimal]	[parent, optimal]
number of permutations	N/A	[3, 5, 10, 15, 20]

Table 4: Characterization of the different data generation settings used in our experiments. There are a total of 5670 simulations for the case with random uncertainty and 3150 simulations for PC uncertainty. Additionally for simulations for more than 9 nodes we only vary the sure edge probability, the forbidden edge probability and the random seed (to 3 values) giving us a total of 27 simulations each for 10 to 25 nodes.

Algorithm 1 Linear additive model (Lagrangian)

```
Input sure edges list, forbidden edges list, index of treatment variable $x$, index of effect variable $y$, number of variables $d$, dataset $D$
Output maximum and minimum of causal query $Q(\hat{A}_\alpha)$
Initialization initialization of $d \times d$ parameter matrix
Loop $k$ iterations
    Sample both soft and hard adjacency matrix $\hat{A}_{\text{hard}}$ and $\hat{A}_{\text{soft}}$ using Gumbel-Softmax, assign the edges to 1 if the edge is in the sure edges list, to 0 if the edge is in the forbidden edges list
    Apply straight-through trick to adjacency matrix $\hat{A}_\alpha \leftarrow (\hat{A}_{\text{hard}} - \hat{A}_{\text{soft}}).\text{detach}() + \hat{A}_{\text{soft}}$
    Get adjustment set $Z(\hat{A}_\alpha, x, y)$ of $x$ on $y$
    Compute $\mathcal{Q}(\hat{A}_\alpha)$ by regressing $D[y]$ on $D[Z]$ and $D[x]$
    Minimize $\mathcal{L}(\hat{A}_\alpha, \lambda, \tau) \leftarrow \pm\mathcal{Q}(\hat{A}_\alpha) + \xi(h(\hat{A}_\alpha), \lambda, \tau)$
```

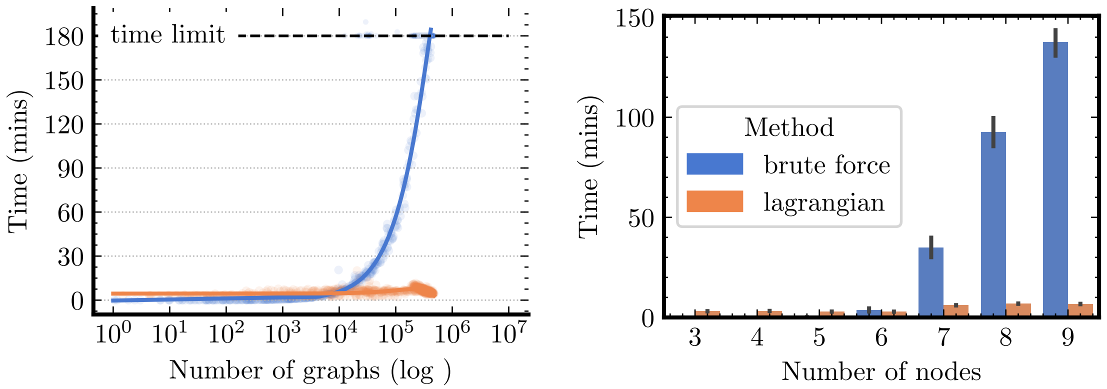
we may falsely consider the ground truth causal query to be out of bounds simply because of finite sample estimation errors. To deal with this, we added an additional step after running the optimization over adjacency matrices. For the final matrices achieving the maximum and minimum values of the causal query, we subsample from the dataset 50 times and estimate the causal query on each of these datasets. Thereby, we obtain a bootstrapped estimation of the variance of our estimated query. In our evaluation, when checking whether the ground truth is within the estimated bounds, we consider the interval $\mathcal{Q} \in [\hat{\mathcal{Q}}_{\min} - \sigma(\hat{\mathcal{Q}}_{\min}), \hat{\mathcal{Q}}_{\max} + \sigma(\hat{\mathcal{Q}}_{\max})]$, i.e., extend the point estimates by one standard deviation of the bootstrapped estimates.

Appendix C. Additional experiments

Fig. 5 shows a relatively steady running time of the non-linear Lagrangian method as the search space grows, next to a super-exponential increase in the computational complexity of the brute force algorithm. On the right, we see this reflected in the number of nodes. Fig. 6 compares running time

Algorithm 2 Non-linear model (Lagrangian)

```
Input sure edges list, forbidden edges list, index of treatment variable $x$, index of effect variable $y$, number of variables $d$, dataset $D$, number of maximum iteration $i == MAX\_ITER$
Output maximum and minimum of causal query $\mathcal{Q}(\hat{A}_\alpha)$
Initialization initialization of $d \times d$ parameter matrix
Loop $k$ iteration
    Sample both soft and hard adjacency matrix $\hat{A}_{\text{hard}}$ and $\hat{A}_{\text{soft}}$ using Gumbel-Softmax, assign the edges to 1 if the edge is in the sure edges list, to 0 if the edge is in the forbidden edges list
    Apply straight-through trick to adjacency matrix $\hat{A}_\alpha \leftarrow (\hat{A}_{\text{hard}} - \hat{A}_{\text{soft}}).\text{detach}() + \hat{A}_{\text{soft}}$
    Get adjustment set $Z(\hat{A}_\alpha, x, y)$ of $x$ on $y$
    Initialize an MLP $f$ with input size of $len(Z) + 1$ and output size of 1
    For $i \leftarrow 1$ to $MAX\_ITER$ do
        Compute $\mathcal{L}(D, Z, x, y) \leftarrow (D[y] - f(D[x, Z]))^2$
        Minimize $\mathcal{L}(D, Z, x, y)$
        if $\mathcal{L}(D, Z, x, y)$ converged or $i == MAX\_ITER$
            break
    Freeze the weights of the MLP $f$
    Compute $\mathcal{Q}(\hat{A}_\alpha) \leftarrow \text{mean}(f(D[Z], D[x] = 1) - f(D[Z], D[x] = 0))$
    Minimize $\mathcal{L}(\hat{A}_\alpha, \lambda, \tau) \leftarrow \pm\mathcal{Q}(\hat{A}_\alpha) + \xi(h(\hat{A}_\alpha), \lambda, \tau)$
```

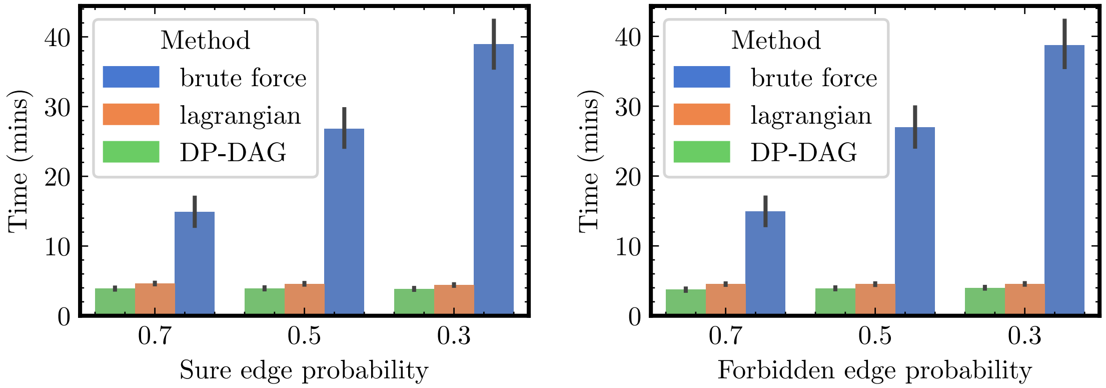


(*a*) Running time comparison for increasing search space size (*3780* simulations)

(*b*) Running time comparison for graphs with different number of nodes (*3780* simulations)

Figure 5: Comparison of running times for the Lagrangian and brute force methods for the non-linear mechanisms.

for different values of sure and forbidden edge probabilities used to generate the set of uncertainties. The full mechanism for this is described in Appendix B.2. As expected, when the sure or forbidden edge probability is low, it implies that a lower proportion of edges from the ground truth are known to be sure or forbidden. This increases the search space and, therefore, the runtime, which is what we see in the plot.



(*a*) Running time comparison for sure edge probability (*5670* simulations)

(*b*) Running time comparison for forbidden edge probability (*5670* simulations)

Figure 6: Comparison of running times for the Lagrangian, DP-DAG and brute force methods for different sure and forbidden edge probabilities used to generate the set of uncertainties.

	Point Coverage		**Bound Coverage**		**Bound Narrowness**	
`Adjustment`	Lagrangian	DP-DAG	Lagrangian	DP-DAG	Lagrangian	DP-DAG
`parent`	0.98 ± 0.0	$\mathbf{1 \pm 0.0}$	0.90 ± 0.0	$\mathbf{0.99 \pm 0.0}$	$\mathbf{2.33 \pm 0.51}$	2.69 ± 0.20
`optimal`	0.98 ± 0.0	$\mathbf{1 \pm 0.0}$	0.92 ± 0.0	$\mathbf{0.99 \pm 0.0}$	$\mathbf{1.99 \pm 0.05}$	2.85 ± 0.22

Table 5: Metric values across different adjustment types for the random uncertainty generation.

Table 5 and Table 6 show the metric values for more configurations. In Table 5, we see that we have effectively perfect point coverage of the ground truth effect value for both types of adjustment sets. However, we see that DP-DAG is better than the Lagrangian at covering the full bound set. In Bound narrowness, we see that the Lagrangian is better again due to the DP-DAG not being able to account for sure edges in its optimization and, therefore, giving wider bounds as expected.

	Point Coverage		**Bound Coverage**		**Bound Narrowness**	
`Mechanism`	Lagrangian	DP-DAG	Lagrangian	DP-DAG	Lagrangian	DP-DAG
`linear`	0.64 ± 0.01	$\mathbf{0.96 \pm 0.01}$	0.63 ± 0.01	$\mathbf{0.76 \pm 0.01}$	$\mathbf{1.25 \pm 0.08}$	1.67 ± 0.09
`non-linear`	0.97 ± 0.0	N/A	0.90 ± 0.0	N/A	2.19 ± 0.05	N/A

Table 6: Metric values for PC-uncertainty. We see that the values for point coverage are lower because it is not guaranteed in this uncertainty generation that the true bounds actually contain the true value, as described in Appendix B.2.

Appendix D. Implementation details

D.1. Adjustment sets

Without properly accounting for confounders, any observed association between the treatment and the outcome may reflect not just the treatment's effect but also the confounders' influence. As a special case, parent adjustment takes the parent set of treatment variable X; it naturally satisfies the definition of an adjustment set, which blocks all potential back-door paths and, therefore, all non-causal paths. Although parent adjustment has the advantage of easy applicability, it has a larger asymptotic variance, which may lead to imprecise causal effect estimation. This becomes even more inefficient when the parent of treatment shows strong co-linearity with the treatment variable. This inefficiency arises because the parent variables may capture too much variability from treatment, which is unnecessary for estimating the total effect on the outcome. Despite the complexity, an optimal adjustment set (Eq. (6)) includes nodes that are not directed cause of treatment X but add precision variable explaining additional variance in the outcome Y, therefore leading to minimum asymptotic estimation variance (Henckel et al., 2022). In Eq. (6), cn stands for causal nodes, which are nodes on the directed path from X to Y, and forbidden nodes $forb$ are defined as the descendent of the causal path and X itself.

$$O(X, Y, \mathcal{G}) = pa(cn(X, Y, \mathcal{G})) \backslash forb(X, Y, \mathcal{G}) \tag{6}$$

Using an adjustment set to compute causal effect includes a selection procedure that is not differentiable. Therefore, we use a matrix multiplication trick that allows the gradient to backpropagate through the desired entries of the adjacency matrix. We summarize our method in Alg. 3.

Algorithm 3 Gradient Preserving Variable Selection

```
Require: Observational dataset data, adjustment set indices s (1D tensor)
Create a selection mask for the adjustment variables: selection_mask $\leftarrow$ s
Initialize an empty list for mask_list
For each index idx in selection_mask (non-zero elements)
    Create a zero matrix mult of size (s.shape[0], s.shape[0])
    Set mult[idx[1], idx[1]] $\leftarrow$ 1
    Append selection_mask $\times$ mult to mask_list
Convert the mask_list to a 2D tensor along axis 0
Compute data_adjustment $\leftarrow$ data $\times$ mask_list$^T$    $\triangleright$ Matrix multiplication to select parent columns
```

When using parent adjustment, there is no ambiguity about which edges are blocked. However, this is not the case for the optimal adjustment. When a variable $\boldsymbol{Z}$ is in the adjustment set, we don't know which edges we need to block (i.e., which entries in the adjacency matrix). $\boldsymbol{Z}$ may have multiple children on the causal path from X to Y. Since estimating the causal effect using an adjustment set does not depend on their children, i.e., edges from the adjustment variable to its children on the causal path are equally good for estimation, we simply take equal contributions from these edges.

D.2. Straight-Through Gumble-Softmax distribution

In our case, the probability of an edge not existing is defined via

$$a_0 = \frac{\exp\left((\log(1-\pi_1)+g_0)/\tau\right)}{\exp\left((\log(1-\pi_1)+g_0)/\tau\right)+\exp\left((\log(\pi_1)+g_1)/\tau\right)}\,, \tag{7}$$

where π_1 is the probability of an edge being 1, g_0 and g_1 are drawn from $\mathrm{Gumbel}(0,1)$ distribution (Gumbel, 1935, 1941), and τ is the temperature parameter. When τ approaches 0, the distribution becomes one-hot. By contrast, when τ becomes large, the distribution tends to be uniform. During optimization, we take a large initial value for τ for better searching and anneal the temperature in order to obtain a static adjacency matrix. Similarly, if an edge exists, the probability can be written as:

$$a_1 = \frac{\exp\left((\log(\pi_1)+g_1)/\tau\right)}{\exp\left((\log(1-\pi_1)+g_0)/\tau\right)+\exp\left((\log(\pi_1)+g_1)/\tau\right)} \tag{8}$$

D.3. Lagrangian optimization

With the notation as defined in the paper, the Lagrangian we aim to minimize with respect to α can be formulated as:

$$\mathcal{L}(\alpha,\lambda,\tau) := \pm\mathcal{Q}(\alpha)+\xi(h(\alpha),\lambda,\tau) \qquad \text{with} \quad \xi(h(\alpha),\lambda,\tau) := -\lambda h(\alpha)+\frac{\tau h(\alpha)^2}{2} \tag{9}$$

where $-/+$ is used for the upper/lower bound. τ is increased throughout the optimization procedure and is seen as a temperature parameter.

Table 7: Overview of resources used in our work.

Name	Reference	License
Python	van Rossum and Drake (2009)	PSF License
PyTorch	Paszke et al. (2019)	BSD-style license
Numpy	Harris et al. (2020)	BSD-style license
Pandas	pandas development team (2020); Wes McKinney (2010)	BSD-style license
Matplotlib	Hunter (2007)	modified PSF
Scikit-learn	Pedregosa et al. (2011)	BSD 3-Clause
SciPy	Virtanen et al. (2020)	BSD 3-Clause
SLURM	Yoo et al. (2003)	modified GNU GPL v2
networkx	Hagberg et al. (2008)	BSD 3-Clause
DCDI	Brouillard et al. (2020)	MIT license
optimaladj	Smucler et al. (2021); Smucler and Rotnitzky (2022)	MIT license

Given an approximate minimum α of this subproblem, we then update λ and τ according to $\lambda \leftarrow \max\{0, \lambda - \tau h(\alpha)\}$ and $\tau \leftarrow \alpha \cdot \tau$ for a fixed $\alpha > 1$. The overall strategy is to iterate between minimizing Eq. (9) and updating λ_l and τ. Separately solving this optimization, once for a minimization problem and once for a maximization problem, would give us the bounds.

Enforcing sure and forbidden edges: For the Lagrangian method, we enforce sure and forbidden edges by blocking the gradients through masking. We construct a masking matrix M of size $d \times d$

with $M_{ij} = 0$ if $(i, j) \in \boldsymbol{E}_s \cup \boldsymbol{E}_f$ and $M_{ij} = 1$ otherwise. Each time a graph is sampled, we mask out the sure and forbidden edges in the adjacency matrix with the mask and replace the sure edges with 1. For the DP-DAG method, only forbidden edges are ensured, therefore, the masking matrix M is defined with $M_{ij} = 0$ if $(i, j) \in \boldsymbol{E}_f$. Similarly, after a DAG is sampled, we apply the mask to its adjacency matrix to remove the forbidden edges.

Appendix E. Resources

Our project heavily relies on available open-source software packages and data sources, which we list in Table 7.