Causal Entropy Optimization

Nicola Branchini
University of Edinburgh
The Alan Turing Institute

Virginia Aglietti
University of Warwick

Neil Dhir
The Alan Turing Institute

Theodoros Damoulas
University of Warwick
The Alan Turing Institute

Abstract

We study the problem of globally optimizing the causal effect on a target variable of an *unknown causal graph* in which interventions can be performed. This problem arises in many areas of science including biology, operations research and healthcare. We propose Causal Entropy Optimization (CEO), a framework which generalizes Causal Bayesian Optimization (CBO) Aglietti et al. (2020b) to account for all sources of uncertainty, including the one arising from the causal graph structure. CEO incorporates the causal structure uncertainty *both* in the surrogate models for the causal effects and in the mechanism used to select interventions via an information-theoretic acquisition function. The resulting algorithm automatically trades-off structure learning and causal effect optimization, while naturally accounting for observation noise. For various synthetic and real-world structural causal models, CEO achieves faster convergence to the global optimum compared with CBO while also learning the graph. Furthermore, our joint approach to structure learning and causal optimization improves upon a sequential, structure-learning-first approach.

1 INTRODUCTION

Causal Bayesian Networks (CBNs) (Pearl, 2009) offer a powerful tool for formulating and testing causal relationships among a set of random variables. Representing a system with a CBN allows one to e.g. estimate causal effects or find interventions optimizing a target node. These tasks often assume, either implicitly or explicitly, *exact* knowledge of the underlying causal graph. Therefore, structure learning (also called "causal discovery") from data (Glymour et al., 2019) has received increasing attention in the last few years. In particular, several studies have taken a Bayesian approach by formulating a prior over graphs and selecting interventions to learn the structure via Bayesian Optimal Experimental Design (BOED, Murphy, 2001; Tong and Koller, 2001; Masegosa and Moral, 2013; Hauser and Bühlmann, 2014; Kocaoglu et al., 2017a; Ness et al., 2017; von Kügelgen et al., 2019; Gamella and Heinze-Deml, 2020; Vowels et al., 2021). Among these studies, a Bayesian Optimization (BO) -based algorithm, targeted at structure learning, has been proposed by von Kügelgen et al. (2019) with the goal of reducing the cost of discovering the true graph. A causal BO-based algorithm (Aglietti et al., 2020b) was also recently developed to identify interventions maximizing a target variable, given a known causal graph – a problem named *causal global optimization*, which we refer to simply as causal optimization. While the two works develop BO algorithms tackling the tasks of structure learning and causal optimization separately, the investigator is often interested in learning optimal actions while not having exact knowledge of the causal relationships among variables. This is the challenging setting we consider in this paper which is the first focusing on solving these two problems jointly.

Example Consider a setting where the investigator aims at finding the level of statin drug or aspirin that should be prescribed to a patient in order to minimize the level of prostate specific antigen (PSA).

While the investigator might have a good understanding of the variables affecting the level of PSA (see Fig. 1), she might not know the exact causal relationships among them. Therefore, multiple causal graphs might be consistent with her domain knowledge. For instance, the causal relationships between age, body mass index (BMI) and cancer might be known, but those among cancer, PSA and different levels of medication administration might be unknown. This is represented by the red edges in Fig. 1. Identifying the optimal drug dosages with CBO would require knowledge of these edges. Indeed, CBO *assumes* (Aglietti et al., 2020b, Appendix, Fig. 3) them to be oriented as $\{\text{Statin} \to \text{Cancer}, \text{Statin} \to \text{PSA}, \text{Aspirin} \to \text{Cancer}, \text{Aspirin} \to \text{PSA}\}$. We propose Causal Entropy Optimization (CEO), a framework which, instead of assuming

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

a specific causal graph for the data generating process, accounts for the structure uncertainty by employing a Bayesian prior. CEO considers the causal graph prior both in the surrogate models and the acquisition function used within the BO algorithm.

Contributions We make the following contributions:

• We generalize the causal global optimization problem to settings where the graph structure is fully or partially unknown.

• We offer the first solution to this problem which we call Causal Entropy Optimization (CEO). CEO models the causal effects via a set of surrogate models that account for both graph and observation uncertainty.

• We introduce an information-theoretic acquisition function, which we call *causal entropy search* (CES) that addresses the trade-off between learning the causal graph and identifying the best intervention. CES encompasses existing acquisitions used for BO and experimental design for structure learning as special cases. It is the first information-theoretic acquisition used in the causal optimization setting.

• We demonstrate across synthetic and real-world causal graphs how accounting for uncertainty leads to faster convergence to the global optimum compared to CBO. In addition, we show how exact knowledge of the causal structure is not always needed for finding the optimal intervention thus a sequential approach learning the causal graph first and then optimizing the target variable might lead to inefficiencies.

2 BACKGROUND AND PROBLEM STATEMENT

We consider a probabilistic causal model (Pearl, 2009) consisting of a directed acyclic graph $\mathcal{G}$ (DAG) and a four-tuple $\langle \mathbf{U}, \mathbf{V}, F, p(\mathbf{U})\rangle$, where $\mathbf{U}$ is a set of *exogenous* background variables distributed according to $p(\mathbf{U})$, $\mathbf{V}$ is a set of observed *endogenous* variables and $F = \{f_1, \dots, f_{|\mathbf{V}|}\}$ is a set of functions constituting the structural causal model (SCM) such that $v_i = f_i(pa_i, u_i)$ with pa_i denoting the parents of V_i. Within $\mathbf{V}$, we distinguish between non-manipulative variables $\mathbf{C}$ that cannot be intervened on, *continuous* treatment variables $\mathbf{X}$ that can be set to specific values and a single output variable Y representing the agent's outcome of interest – see Fig. 1 for an example. Under the *Markov Assumption*, each variable V_i is conditionally independent of its non-descendants given its parents, such that the joint distribution factorises as: $p(\mathbf{V} \mid \mathcal{G}) = \prod_{V_j \in \mathbf{V}} p(V_j \mid \mathbf{Pa}_j^{\mathcal{G}}, \mathcal{G})$. Denote by $\mathcal{P}(\mathbf{X})$ the power set of $\mathbf{X}$ giving *all* possible interventions we can perform in the graph. The set $\mathbf{X}_I \in \mathcal{P}(\mathbf{X})$ represents one such intervention set with $\mathbf{V}_I = \mathbf{V} \setminus \mathbf{X}_I$ being the corresponding set of non-intervened variables. We assume *causal sufficiency* (Sgouritsa, 2015) and *perfect* interventions (Peters et al., 2017) so that, for any set $\mathbf{X}_I$, the interventional distribution $p(\mathbf{V}_I \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I))$ resulting from setting $\mathbf{X}_I$ to a value $\mathbf{x}_I$ obeys the following *truncated factorization*

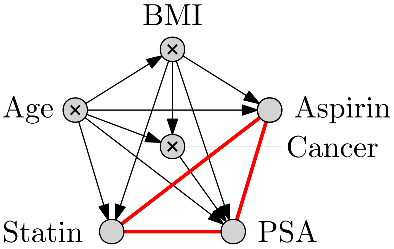


Figure 1: Causal DAG with unoriented edges in red. Shaded and crossed nodes represent manipulative and non-manipulative variables respectively. PSA is the outcome of interest.

$$p(\mathbf{V}_I \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}) = \prod_{V_j \in \mathbf{V}_I} p(V_j \mid \mathbf{Pa}_j^{\mathcal{G}})\Big|_{\mathbf{X}_I = \mathbf{x}_I}$$

where the vertical line represents evaluation of the expression at $\mathbf{X}_I = \mathbf{x}_I$.

Notation Note that, throughout, lowercase denotes *realizations* of random variables (r.v.), while uppercase denotes the r.v. We denote collected data by $\mathcal{D} = \{\{(\mathbf{x}_I^{(i)}, \mathbf{v}_I^{(i)})\}_{i=1}^N\}_{\mathbf{X}_I \in \mathbf{ES}}$ where the *exploration set* (**ES**) is $\mathbf{ES} = \mathcal{P}(\mathbf{X})$ or $\mathbf{ES} \subseteq \mathcal{P}(\mathbf{X})$ if only a subset of interventions can be implemented in the system (e.g. if $\mathcal{P}(\mathbf{X})$ is too large). Here, $\mathbf{v}_I^{(i)}$ represents one set of values of the *non-intervened variables* $\mathbf{V}_I$ in a mutilated graph where $\mathbf{X}_I$ is set to $\mathbf{x}_I^{(i)}$. Further, $\mathbf{v}_I^{(i)}$ includes both a value for the target variable $y^{(i)}$ and a value for the remaining variables, denoted by $\mathbf{v}_{I_Y}^{(i)} = \mathbf{v}_I^{(i)} \setminus y^{(i)}$. The dataset $\mathcal{D}$ includes all data, that is observational data ($\mathcal{D}^O$) which correspond to $I = \varnothing$ and interventional data ($\mathcal{D}^I$). Every time we intervene in the system we collect $N > 1$ samples from each interventional distribution, but the framework equally handles $N = 1$. See Appendix A for a table describing the full notation.

Problem statement We consider the causal global optimization problem introduced by Aglietti et al. (2020b) and generalize it to settings with an *unknown* causal graph. Given $\mathcal{D}$, we seek to identify the interventional set $\mathbf{X}_I$ *and* corresponding values $\mathbf{x}_I$ optimizing the causal effect in the *true* causal graph $\mathcal{G}$

$$\mathbf{X}_I^\star, \mathbf{x}_I^\star = \mathop{\arg\min}_{\substack{\mathbf{X}_I \in \mathcal{P}(\mathbf{X}) \\ \mathbf{x}_I \in D(\mathbf{X}_I)}} \mathbb{E}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}], \qquad (1)$$

where $D(\mathbf{X}_I)$ denotes the interventional domain of $\mathbf{X}_I$ and the causal effect depends on the true graph $\mathcal{G}$. Solving the problem in Eq. (1) is challenging as evaluating $\mathbb{E}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}]$ requires intervening in the real system at a *cost*, which we assume to be given by $\text{Co}(\mathbf{X}_I, \mathbf{x}_I)$. *While Eq.* (1) *is the objective of* CBO*, since the graph is unknown the problem becomes more challenging.*

Remark: Notice that, when the graph is known, one can solve the problem in Eq. (1) resorting to CBO. *When the graph is unknown, there is no existing unified solution for the problem.* This paper offers such a solution and shows how, accounting for graph uncertainty, one can match the performance of CBO, which requires knowledge of the true graph, in terms of convergence speed to the optimum.

3 METHODOLOGY

There are *three* main ingredients to our solution to the problem in Eq. (1): a set of surrogate models for the causal effects on the target node (§3.1) accounting for different sources of uncertainty, a posterior over the graph structure (§3.2), computed exploiting observational and interventional data, and an information-theoretic acquisition function (§3.3) balancing the trade-off between optimization and structure learning. We now detail each ingredient and provide the pseudocode for CEO in Algorithm 1.

3.1 Inference on the causal effects

Prior Surrogate Models For each set $\mathbf{X}_I \in \mathbf{ES}$ we place a Gaussian process (GP, Williams and Rasmussen, 2006) prior on $g_I(\mathbf{x}_I) = \mathbb{E}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}]$ and construct a prior mean and kernel function that incorporates our current belief about the graph together with the observational and interventional data. Specifically, we define $g_I(\mathbf{x}_I) \sim \mathcal{GP}(m_I(\mathbf{x}_I), k_I(\mathbf{x}_I, \mathbf{x}'_I))$, with

$$m_I(\mathbf{x}_I) \overset{\text{def}}{=} \widehat{\mathbb{E}}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I)] \tag{2}$$

$$k_I(\mathbf{x}_I, \mathbf{x}'_I) \overset{\text{def}}{=} k(\mathbf{x}_I, \mathbf{x}'_I) + \widehat{\mathbb{V}}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I)], \tag{3}$$

where $k(\cdot, \cdot)$ is a problem-dependent kernel function of choice. We incorporate the uncertainty over the graph by introducing latent variable G, by defining

$$\widehat{\mathbb{E}}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I)] \overset{\text{def}}{=} \mathbb{E}\Big[\widehat{\mathbb{E}}[Y|\text{do}(\mathbf{X}_I = \mathbf{x}_I), G]\Big] \tag{4}$$

$$\widehat{\mathbb{V}}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I)] \overset{\text{def}}{=} \mathbb{E}\Big[\widehat{\mathbb{V}}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), G]\Big] + \mathbb{V}[\widehat{\mathbb{E}}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), G]],$$

where *outer* expectations/variances are w.r.t a probability mass function on the r.v. G, i.e. $P(G)$, whereas *inner* expectations/variances are w.r.t the approximation $\widehat{p}(Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), G = g)$, a Monte Carlo estimate of an interventional distribution computed via the do-calculus with only observational data [1]. Note that, when collecting data, $P(G)$ will be replaced by $P(G \mid \mathcal{D})$, which also includes interventions. Therefore, our surrogate models combine observational and interventional data. Computing the posterior $P(G|\mathcal{D})$ will be discussed in detail in §3.2.

Surrogate Model Likelihood For each intervention set $\mathbf{X}_I \in \mathbf{ES}$ and value $\mathbf{x}_I^{(i)}$, the output y is a *noisy* realisation of the objective function $y = g_I(\mathbf{x}_I) + \nu$ at $\mathbf{x}_I^{(i)}$ where $\nu \sim \mathcal{N}(0, s^2)$. Every time we perform $\text{do}(\mathbf{X}_I = \mathbf{x}_I^{(i)})$, we obtain a single sample $\mathbf{v}_I^{(i)}$ from the resulting interventional distribution $p(\mathbf{V}_I|\text{do}(\mathbf{X}_I = \mathbf{x}_I^{(i)}), \mathcal{G})$. Noisy observations were not considered by CBO, which made a simplifying assumption, and complicate the identification of the optimal intervention.

Posterior Surrogate Models Given the Gaussian likelihood, the posterior distribution $p(g_I \mid \mathcal{D})$ can be derived analytically and will also be a GP with mean $m_I(\mathbf{X} \mid \mathcal{D})$ and covariance functions $k_I(\mathbf{x}_I, \mathbf{x}'_I \mid \mathcal{D})$ computed by standard GP updates (Williams and Rasmussen, 2006, p. 19).

3.2 Inference on the causal graph

Given $\mathcal{D}$, we update the prior distribution on G to get $P(G \mid \mathcal{D})$. Our updated uncertainty on the graph structure is then reflected in the surrogate models. We follow the setting of von Kügelgen et al. (2019) and consider a discrete uniform prior distribution on G with $P(G = g) = \frac{1}{|R_G|}$. Notice that, given this prior formulation, we assume to be able to enumerate all potential causal graphs. This is appropriate for intended applications of CBO, where a manageable number of graphs representing alternative causal hypotheses is maintained. This allows us to compute expectations under $P(G|\mathcal{D})$ without approximations. However, in our framework only *expectations* w.r.t. the graph posterior are required, and can thus in principle be approximated.

Graph Likelihood In order to define the graph likelihood, we assume an additive noise model with i.i.d. Gaussian noise terms[2]. For every g and $X_j \in \mathbf{X}$, we have

$$X_j = f_j(\mathbf{Pa}_j^{G=g}) + u_j, \tag{5}$$

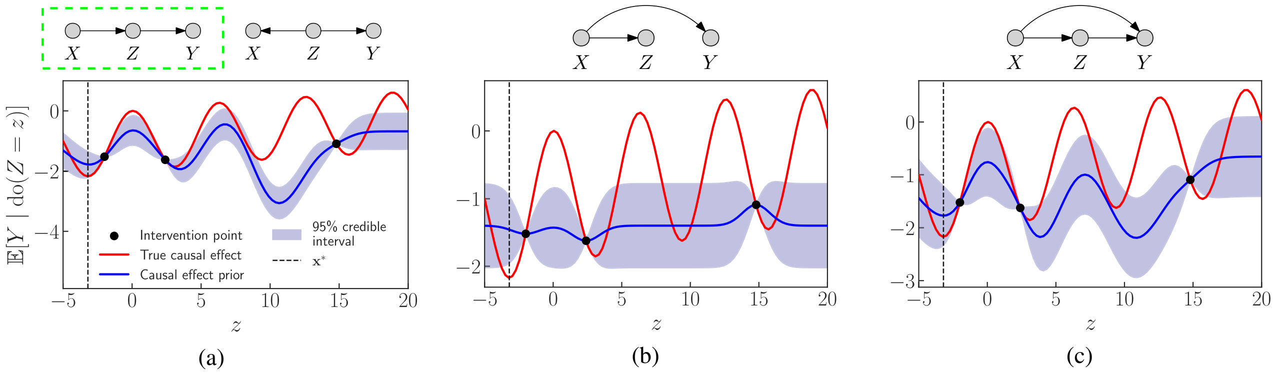
$$f_j \mid (\mathbf{Pa}_j^{G=g} = \mathbf{x}, \theta_j) \sim \text{GP}(0, k_j(\mathbf{x}, \mathbf{x}'; \theta_j)) \tag{6}$$

with $u_j \sim \mathcal{N}(0, \sigma_j^2)$ and where θ_j represents a set of hyperparameters. Here, $\mathbf{x}$ and $\mathbf{x}'$ are two values for the parents of V_j in $G = g$ and k_j is a kernel function of choice. We assume parameter modularity (Friedman and Nachman, 2000), i.e. the conditionals in Eq. (6) depend on the choice of G only through the choice of parents, not on other aspects of G. Exploiting the available data we can then compute $p(f_j \mid \mathbf{Pa}_j^G, \mathcal{D}^O)$ in closed form, due to the additive noise. The graph likelihood is given by the product of GP marginal likelihoods, corresponding to the truncated factorization of §2 (details on likelihood computation in Appendix B.)

Graph Posterior Given the prior distribution on G and the likelihood, we can compute the posterior distribution on G *in closed form* as $P(G \mid \mathcal{D}) \propto p(\mathcal{D}|G)P(G)$ due to the

[1] This can be computed when the causal effect is identifiable (Pearl, 2009). See Appendix B for a discussion of the conditions under which Eqs. 4 converge to the true $\mathbb{E}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}]$ and $\mathbb{V}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}]$ respectively.

[2] This is a common assumption, see e.g. von Kügelgen et al. (2019) or Silva and Gramacy (2010).



(a)

(b)

(c)

Figure 2: Each figure plots the true underlying causal effect function $\mathbb{E}[Y \mid \mathrm{do}(Z = z), G = \mathcal{G}]$ in red, while showing different surrogate models arising from Eqs. 4 when assuming each of the graphs of top of the plot is the correct one. The surrogate models share hyper-parameters so the differences only stems from alternative causal assumptions. The true DAG $\mathcal{G}$ is shown inside the dashed box in Fig. 2a. Vertical dashed line indicates the global optimum. Note that both the true graph and the second DAG above (a) result in the same surrogate model.

tractability of GPs. Importantly, this also allows us to compute the expectations w.r.t. $P(G|\mathcal{D})$ without approximations in Eqs. 4 and later in the acquisition function.

Convergence Under some conditions, convergence to $\delta_{G=\mathcal{G}}$ can be guaranteed asymptotically (in the number of interventional and observational samples). The key requirements needed are: causal sufficiency; R_G including $\mathcal{G}$; all variables can be manipulated; infinite samples can be obtained from each node; causal minimality. See Appendix B for details.

Effect of graph knowledge on optimization It is important to emphasize that while computing $P(G \mid \mathcal{D})$ and using it in the surrogate model provides proper uncertainty quantification, exact knowledge of the causal structure is not *always* needed for more efficient causal optimization.

Indeed, different graphs may lead to equivalent surrogate models. For instance, consider the example in Fig. 2 showing different causal graphs and the associated surrogate models, when three interventions are collected. Here the true DAG is given in Fig. 2a (green dashed box) alongside an alternate causal DAG. The surrogate models for these structures are equivalent, and shown in the plot below. Indeed, the do-calculus for the left DAG and right DAG of Fig. 2a gives the same result, even if the latter DAG is wrong. Given our surrogate model construction in (Eq. (2)), this implies the same prior mean and covariance for the causal effects associated with both graphs. Therefore, for the purpose of causal optimization, the DAGs in Fig. 2a are the same.

As demonstrated experimentally in §5, *it is thus wasteful to intervene to find the true graph first, and only after perform causal optimization*. This motivates our joint approach, which automatically balances structure learning and optimization by picking interventions that are reducing the uncertainty of $P(G \mid \mathcal{D})$ *when* doing so enables faster identification of the optimum. Next, we describe how to achieve this balance by proposing a new acquisition function.

3.3 Acquisition function: Causal Entropy Search

We propose an acquisition function that combines Bayesian structure learning and causal optimization into a single objective in order to balance the two tasks. To do so, we start by framing the tasks of causal optimization and structure learning, respectively, as *gaining information* about the global optimum *value* $y^\star = \min(\text{or} \max)_I y_I^\star$ and about G. This information gain can be quantified by considering the reduction in uncertainty in an appropriately defined joint distribution on $y^\star$ and G. A distribution on the graph, $P(G \mid \mathcal{D})$, was defined in §3.2 and the distribution on $y^\star$ is implicitly *induced* by the surrogate models in §3.1 [3]. Exploiting the distribution on $y_I^\star$ (i.e. within a single surrogate model) has precedents in BO with *output-space* entropy search methods (Garnett, 2022, Chapter 8, Section 8.9).

Our acquisition is thereby defined as the conditional mutual information (MI) denoted by $\mathbb{I}(\cdot;\cdot \mid \cdot)$ between the random variables $(y^\star, G)$ and the *outcome of the experiment* $(\mathbf{v}_Y, y)$, given data $\mathcal{D}$. An experiment in the context of a causal DAG consists of performing $\mathrm{do}(\mathbf{X}_I = \mathbf{x}_I)$ and observing a sample $\mathbf{v}_I = (y, \mathbf{v}_Y)$ from the resulting interventional distribution. Therefore, in CEO we seek an intervention set $\mathbf{X}_I^{\text{best}}$ *and* corresponding value $\mathbf{x}_I^{\text{best}}$ such that

$$\mathbf{X}_I^{\text{best}}, \mathbf{x}_I^{\text{best}} = \operatorname*{arg\,max}_{\mathbf{X}_I \in \mathbf{ES},\ \mathbf{x}_I \in D(\mathbf{X}_I)} \alpha_{\text{CES}}(\mathbf{X}_I, \mathbf{x}_I)$$

$$\alpha_{\text{CES}}(\mathbf{X}_I, \mathbf{x}_I) = \frac{\mathbb{I}\left[(y^\star, G)\,;(\mathbf{X}_I, \mathbf{x}_I, \mathbf{v}_Y, y) \mid \mathcal{D}\right]}{\text{Co}(\mathbf{X}_I, \mathbf{x}_I)} \qquad (7)$$

[3] This is because the distribution on each $y_I^\star$, induced by the prior on f_I, implies a distribution on the global optimum value achieved $y^\star$. Formally, each $y_I^\star$ has the distribution of a minimum/maximum of a Gaussian process, which has been studied (Adler et al., 2007).

Algorithm 1 CES

```
Input: $\mathcal{D}$, $P(G)$, $H$ (N. of iterations)
Output: $\mathbf{X}_I^\star, \mathbf{x}_I^\star, P(G \mid \mathcal{D}_H)$.
Initialise: Set $\mathcal{D}_0 = \mathcal{D}$.
Compute $p(\mathcal{D}_0 \mid G = g)$ for each $g \in R_G$.
Compute $P(G \mid \mathcal{D}_0)$.
Compute $m_I(\mathbf{x}_I)$ and $k_I(\mathbf{x}_I, \mathbf{x}'_I)$ using the do-calculus for each $\mathbf{X}_I \in \mathbf{ES}$ (Eqs. 4 ).
for $h = 1, ..., H$ do
    Compute $\alpha_{\text{CES}}(\mathbf{X}_I, \mathbf{x}_I)$ for each $\mathbf{X}_I \in \mathbf{ES}$ and each $\mathbf{x}_I$ in the acquisition points with CES (Algorithm 2.
    Obtain the optimal set-value pair $(\mathbf{X}_I^h, \mathbf{x}_I^h)$.
    Intervene on the system and augment the dataset $\mathcal{D}_h = \mathcal{D}_{h-1} \cup (\mathbf{x}_I^h, \mathbf{v}_I^h)$.
    Compute $P(G \mid \mathcal{D}_h)$ and every $p(g_I \mid \mathcal{D}_h)$.
end for
Return $(\mathbf{X}_I^\star, \mathbf{x}_I^\star), P(G|\mathcal{D}_H)$
```

where the denominator $\text{Co}(\mathbf{X}_I, \mathbf{x}_I)$ provides the MI per unit of cost. We call the acquisition function in Eq. (7) Causal Entropy Search (CES).

Assume, for the moment, that we have access to $p(y^\star, G)$ and the associated posterior. We write its joint entropy $\mathbb{H}[p(y^\star, G \mid \mathcal{D})]$ as

$$-\sum_G \int \mathrm{d}y^\star p(G, y^\star \mid \mathcal{D}) \log p(G, y^\star \mid \mathcal{D}). \qquad (8)$$

The work in (Marx et al., 2021, Lemma 2.3 and Eq. 2.2) shows that conditional MI and joint entropies can be rigorously defined in an analogue way to their fully continuous/discrete counterparts. We can then use Eq. (8) to evaluate the numerator of CES, i.e. Eq. (7), which can be written

$$\mathbb{E}\Big[\mathbb{H}[p(y^\star, G \mid \mathcal{D})] - \mathbb{H}[p(y^\star, G \mid \mathcal{D} \cup (\mathbf{x}_I, \mathbf{v}_Y, y))]\Big],$$

where the expectation is w.r.t $p(\mathbf{v}_Y, y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{D})$. This is the reduction in entropy observed *on average* in the joint distribution on $(y^\star, G)$ when performing $\text{do}(\mathbf{X}_I = \mathbf{x}_I)$ and observing $\mathbf{v}_Y = (\mathbf{v}_Y, y)$. We note that computing this term does not require intervening in the real system. For a given intervention we “fantasize” (a concept and terminology often used in BO (Wilson et al., 2018)) about the outcomes $\mathbf{v}_Y$ we would get by simulating what would happen in the system under $\text{do}(\mathbf{X}_I = \mathbf{x}_I)$. This can be done in CEO by using the surrogate models and the fitted SCM functions. Note that this formulation automatically takes into account noisy observations, which often motivates entropy-based acquisitions in BO (Frazier, 2018). We now discuss how to define and obtain $p(y^\star, G)$, and approximate $\mathbb{H}[p(y^\star, G)]$.

Joint posterior over $y^\star$ and G Computing CES requires defining $p(y^\star, G)$, obtaining its posterior given $\mathcal{D}$ and computing Eq. (8). We model the joint distribution as $p(y^\star, G \mid \mathcal{D}) = P(G \mid y^\star, \mathcal{D}) p(y^\star|\mathcal{D})$[4]. We can thus write Eq. (8) as the sum of: $-\sum_G \int \mathrm{d}y^\star P(G \mid y^\star, \mathcal{D}) p(y^\star \mid \mathcal{D}) \log P(G \mid y^\star, \mathcal{D})$ and $\mathbb{H}(p(y^\star \mid \mathcal{D}))$. In turn, evaluating these two terms require computing $P(G \mid y^\star, \mathcal{D})$ and sampling from $p(y^\star \mid \mathcal{D})$. The former can be achieved by updating the posterior on G to get $P(G \mid \mathcal{D} \cup (\mathbf{x}^\star, \mathbf{v}_Y^\star, y^\star))$ and marginalizing over $\mathbf{x}^\star$ and $\mathbf{v}_Y^\star$. However, since $p(y^\star \mid \mathcal{D})$ is not in available closed form, and a Monte Carlo approximation is computationally expensive, we propose a simple approximation that works in practice. Specifically, we approximate the distribution of $y^\star$ to be a *mixture* of the intervention-specific distributions on the optimal targets, where the weights are given by the probability of each intervention set being optimal

$$p(y^\star \mid \mathcal{D}) = \sum_{y_I^\star : \mathbf{X}_I \in \mathbf{ES}} P(\mathbf{X}_I = \mathbf{X}^\star)\, p(y_I^\star \mid \mathcal{D}), \qquad (9)$$

where $\mathbf{X}^\star$ is the intervention set in $\mathbf{ES}$ which yields the global optimum. For example, if we have three surrogate models corresponding to $\mathbf{ES} = \{\{X\}, \{Z\}, \{X, Z\}\}$, and the the global minimum can be found within the causal effect of Z, i.e. $\mathbb{E}[Y|\text{do}(Z = z)]$, then $\mathbf{X}^\star = \{Z\}$.
We can thus sample from Eq. (9) combining standard mixture sampling techniques (Owen, 2013) with sampling from each $p(y_I^\star \mid \mathcal{D})$ as in output-space ES methods (Wang and Jegelka, 2017). Notice that Eq. (9) is accounting for the fact that we do not know which intervention set and thus surrogate model is associated with the global optimum[5]. See Table 1 for connections between the CES and related existing objectives in experimental design and Bayesian optimimization.

Motivation behind posterior approximation Notice that as we collect data and get closer to the optimum, we expect the weights of the mixture distribution of Eq. (9) to gradually concentrate around the optimal intervention set $\mathbf{X}^\star$, and $p(y^\star \mid \mathcal{D})$ to turn into $p(y_I^\star \mid \mathcal{D})$ for the $\mathbf{X}_I$ s.t. $\mathbf{X}_I = \mathbf{X}^\star$. To model the belief over the optimal set $P(\mathbf{X}_I = \mathbf{X}^\star)$, we employ a *multi-armed bandit* (Lattimore and Szepesvári, 2020) perspective and define the weights via an upper confidence bound (UCB) policy; see Appendix G for details.

Computational complexity Let N be the number of *acquisition points* i.e., given $\mathbf{X}_I$, we will consider $\{\mathbf{x}_I^{(n)}\}_{n=1}^N$ as candidate values to compare with CES. The total complexity of CES can be written: $\mathcal{O}(N \cdot \text{CES}(\mathbf{x}) \cdot \sum_{\mathbf{X}_I \in \mathbf{ES}} |\mathbf{X}_I|)$. Here, $\text{CES}(\mathbf{x})$ denotes the time needed to compute CES for a *specific value* $\mathbf{x}$, regardless of its corresponding $\mathbf{X}_I$. Note this is valid for CBO also, replacing $\text{CES}(\mathbf{x})$ with $\text{CEI}(\mathbf{x})$, i.e. the acquisition used by CBO (Aglietti et al., 2020b).

[4] The alternative model $p(y^\star|G, \mathcal{D})P(G|\mathcal{D})$ could also be considered. We discuss this in Appendix F

[5] Therefore, this formulation provides a useful acquisition that could be used not only for causal optimization, but also in a-causal settings.

Here, CES$(\mathbf{x})$ involves approximating univariate marginals $p(y_I^\star \mid \mathcal{D})$ and the mixture in Eq. (9). We do this with Kernel Density Estimation (KDE) whose complexity depends on the accuracy needed to estimate these marginals. Finally, since the number of graphs in our setting is tractable, we can compute expectations w.r.t $P(G)$ without approximations and with negligible cost w.r.t to the rest.

4 RELATED WORK

Causal Effect Estimation. Many approaches to estimate causal effects from observational data have been proposed, including those based on propensity scores (Rosenbaum and Rubin, 1983), instrumental variables (Angrist and Imbens, 1995) or SCMs (Pearl, 2009). Instead, there have been few methods combining interventional and observational data (Silva, 2016). Focusing on SCM methods, apart from a few exceptions (Hyttinen et al., 2015; Horii, 2021), causal effects are estimated assuming *exact* knowledge of $\mathcal{G}$.

Causal Discovery. The majority of causal discovery (CD) methods focus on learning $\mathcal{G}$ using only observational data thus restricting the identification to the Markov equivalence class (MEC) (Verma and Pearl, 1991; Andersson et al., 1997; Spirtes et al., 2000a; Chickering, 2002; Friedman and Koller, 2003; Shimizu et al., 2006; Janzing et al., 2012; Zhang et al., 2015). The seminal work by Cooper and Yoo (1999) first showed how experimental design can improve causal structure learning (which in general is known to be *NP-hard* (Chickering, 1996)). Since this study, several papers have focused on learning $\mathcal{G}$ from a combination of interventional *and* observational data (Tong and Koller, 2001; Murphy, 2001; Eaton and Murphy, 2007b; Hauser and Bühlmann, 2012, 2014, 2015; Wang et al., 2017; Ness et al., 2017; Yang et al., 2018; Ghassami et al., 2018; Agrawal et al., 2019; Faria et al., 2022). However, all of these focus on finding the true graph by selecting the intervention *set* only. Our work additionally selects intervention *values*. Recently, von Kügelgen et al. (2019) developed a BO framework for CD where variables are continuous and follow flexible non-linear relationships.

Optimal Causal Decision Making. The literature on causal decision making has mainly focused on finding the optimal treatment regime using observational data (Zhang et al., 2012; Atan et al., 2018; Håkansson et al., 2020). The idea of identifying the optimal action or policy by performing interventions in a causal system has been explored in causal bandits (Lattimore et al., 2016), causal reinforcement learning Zhang (2020) and, more recently, in BO (Aglietti et al., 2020b,a, 2021). Importantly, all these approaches assume exact knowledge of the causal relationships beforehand, an assumption that is often not met in practice. Recent work attempts to relax this for causal bandits (Lu et al., 2021; Wang and Zhou, 2021).

Table 1: Causal Entropy Search compared to related objectives in BOED for structure learning and BO algorithms. Entropy Search (ES) and Max-value ES are used within the standard a-causal BO problem. Multi-Task BO is also used for a-causal optimization with multiple functions, but when the function containing the optimum is known. The objective in (von Kügelgen et al., 2019) is also based on MI, but only targets the graph structure. CES considers intervention sets, intervention values and the graph structure when the function containing the optimum is unknown.

Acquisition	**Objective**
CES (ours)	$\frac{\mathbb{I}[(y^\star,G);(\mathbf{X}_I,\mathbf{x}_I,\mathbf{v}_Y,y)\mid\mathcal{D}]}{\textsc{Co}(\mathbf{X}_I,\mathbf{x}_I)}$
Entropy Search (Hennig and Schuler, 2012)	$\mathbb{I}\left[(\mathbf{x}, y); \mathbf{x}^\star \middle\vert\, \mathcal{D}\right]$
Max-value Entropy Search (Wang et al., 2017)	$\mathbb{I}\left[(\mathbf{x}, y); y^\star \middle\vert\, \mathcal{D}\right]$
Multi-Task BO (Swersky et al., 2013)	$\frac{\mathbb{I}\left[(\mathbf{x}^{\text{task}},y);\mathbf{x}^\star \middle\vert \mathcal{D}\right]}{\textsc{Co}(\text{task})}$
Causal discovery via BO (von Kügelgen et al., 2019)	$\mathbb{I}\left[G; (\mathbf{X}_I, \mathbf{x}_I, \mathbf{v}_I) \mid \mathcal{D}\right]$

Additional related works. Toth et al. (2022) recently proposed a framework where posterior distributions over both the causal graph and the interventional distributions are learned, with the goal of solving an "integrated causal discovery and reasoning" task. A similar task is solved by Tigas et al. (2022) which exploits submodularity to obtain theoretical guarantees in terms of convergence to the true interventional distributions. Differently, in this paper, we learn a *joint* posterior over the graph and the optimal intervention, as opposed to learning the full set of interventional distributions. A similar difference between finding the optimum function value or learning the full function exists between Bayesian quadrature and Bayesian optimization (Garnett, 2022; Hennig et al., 2022).

5 EXPERIMENTS

We demonstrate CEO on a benchmark synthetic example used in (Aglietti et al., 2020b) as well as real-world applications for which a DAG is available and can be used as a simulator. Without loss of generality, we always *minimize* causal effects rather than maximize. Results are averaged over 12 replicates of different initial $\mathcal{D}^I$, while $\mathcal{D}^O$ is fixed. We set $|\mathcal{D}_0^I| = 2$ and $|\mathcal{D}_0^O| = 200$ (initial data) unless otherwise specified[6]. Full experimental details including SCM details, kernel functions used, hyper-parameter optimization can be found in Section K of the Appendix. As a reminder on notation, recall $\mathcal{G}$ denotes the true causal graph; $y^\star$ denotes the best value of the causal effect across ES.

[6] Code available at: `https://github.com/nicola144/CEO`

Table 2: Average GAP ± one standard error computed across 12 replicates initialized with different $\mathcal{D}$. Higher values are better, and $0 \leq$ GAP ≤ 1. The best result for each experiment in bold.

Method	Synthetic (Fig. 2a)	Health (Fig. 1)	Epi (Fig. 3)	(Epi$^+$ Fig. 9)
CEO	$\mathbf{0.75} \pm 0.03$	0.62 ± 0.07	$\mathbf{0.59} \pm 0.06$	$\mathbf{0.65} \pm \mathbf{0.03}$
CBO w. $G = \mathcal{G}$	0.66 ± 0.04	$\mathbf{0.71} \pm 0.05$	0.50 ± 0.05	0.50 ± 0.03
CBO w. $G \neq \mathcal{G}$	0.59 ± 0.03	0.62 ± 0.06	0.50 ± 0.06	0.53 ± 0.03

Experiments Roadmap & Baselines We compare the performance of CEO against CBO which is the state-of-the-art algorithm addressing the causal global optimization problem. However, as CBO assumes knowledge of the graph, we run it first assuming the true graph $\mathcal{G}$ and then assuming each of the wrong graphs. In general, CBO with $\mathcal{G}$ is expected to perform better, especially when the graph structure is particularly informative for the CBO prior. However, this need not be the case in all experiments. Indeed, in one of our real examples, CBO equipped with $\mathcal{G}$ performs worse than CEO. To further highlight the benefit of our joint method, in §5.2 we compare against an algorithm that first identifies the causal graph by optimizing the MI[7] as considered by von Kügelgen et al. (2019), and then runs CBO to select the optimal action retaining the interventional data collected when learning about the graph. This is to show that jointly considering structure learning and causal optimization leads to better performance than a sequential approach. We refer to this method as CD-CBO. Notice that we cannot compare directly to von Kügelgen et al. (2019) as in their work the MI is optimized via BO rather than computed as in CEO. Finally, notice also that causal discovery methods which do not (1) collect interventions with active learning (2) select both intervention sets and values and (3) assume continuous variables, are not applicable to the problem setting thus we do not consider them as an alternative to von Kügelgen et al. (2019) for CD-CBO.

Performance measures We evaluate performance by assessing the convergence speed to the optimum value of Y as measured by the total *cumulative cost* of interventions taken where the cost of a single intervention is given by the number of variables in the intervened set. Further, we also evaluate our approach using the GAP metric introduced in Aglietti et al. (2021, Eq. (4)), which is defined by

$$\left[\frac{y(\mathbf{x}_I^{\text{best}}) - y(\mathbf{x}_{\text{init}})}{y^\star - y(\mathbf{x}_{\text{init}})} + \frac{H - H(\mathbf{x}_I^{\text{best}})}{H}\right] \Big/ \left(1 + \frac{H-1}{H}\right).$$

Here, $y^\star$ is the true global optimum, $y(\mathbf{x}_I^{\text{best}})$ is the best value achieved by the algorithm, $y(\mathbf{x}_{\text{init}})$ is the value obtained at the beginning of the optimization, H is the total number of iterations, $H(\mathbf{x}_I^{\text{best}})$ is the number of iterations to reach the best value achieved. Note that $0 \leq$ GAP ≤ 1 and that higher values are better.

5.1 Synthetic example

We start by considering data generated by the chain graph $X \to Z \to Y$ as per Fig. 2a (green dashed box). We consider *all* possible alternative DAGs with three nodes as alternative causal hypotheses, except for those which do not make sense for causal optimization: graphs with any isolated nodes or where the target Y is not a sink. See Appendix H for the full list of graphs and the true SCM.

Convergence results are given in Fig. 4, while we show the evolution of the posterior on the true graph over iterations in Appendix D. Note how CBO run under incorrect causal assumptions does not converge to the global optimum, whereas CEO matches the performance of CBO run with the true graph even if the graph posterior has not converged. This is confirmed by the GAP values in Table 2 where CEO outperforms both CBO algorithms across 12 replicates.

5.2 Comparison with learning the structure first, then performing CBO (CD-CBO)

We now study how CEO compares against an algorithm that first performs causal discovery and, once the graph has been identified, solves the optimization problem via CBO. We refer to this as CD-CBO. We consider the graph learned once it has more than 90% of posterior mass. In Fig. 4(a) , notice the significantly worse performance at optimization of CD-CBO. This is due to the fact that, while the MI correctly identifies $\mathcal{G}$ after a few samples for most graphs, the graph in Fig. 2c is very hard to distinguish from $\mathcal{G}$ (for the given SCM) as the terms in the truncated factorization give similar likelihood values. Therefore, the posterior never gets concentrated around a single graph despite the high number of selected interventions, and the optimization task is never solved. This reflects the benefit of a joint approach, where the graph is learned along optimization of the effect, and only to the extent to which it is useful for optimization. Additional results are presented in Appendix M.

[7] We note also note that (Agrawal et al., 2019) use a MI criteria for causal discovery via experimentation. However, this work only selects intervention set and not values and it is not not applicable to our continuous variables setting.

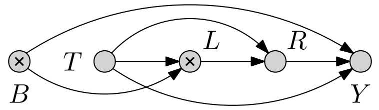
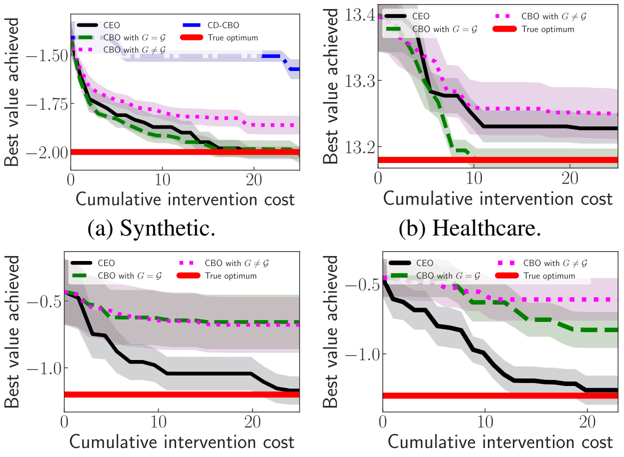


Figure 3: True DAG used for the epidemiology experiment. For the full support of $P(G)$, and SCM equations, see Appendix I Crossed variables are non-manipulative.



(a) Synthetic.

(b) Healthcare.

(c) Epidemiology.

(d) Ext. Epidemiology

Figure 4: Convergence plot for CEO and competitor methods across replicates. In the Healthcare and Epidemiology related examples, the graph is found immediately by the posterior with the initial $\mathcal{D}^O$ and $\mathcal{D}^I$; therefore CD-CBO *simply reduces to* CBO. Shaded areas are ± one standard error. Vertical axis shows best value of the causal effect achieved, horizontal shows *cumulative* intervention cost.

5.3 Real examples

We now study the performance of CEO with causal DAGs used in two real-world settings: one in healthcare to model the level of Prostate Specific Antigen (PSA) and one in epidemiology to model the level of HIV virus load. These graphs exhibit a significantly more complex dependency structure than the chain graph of the synthetic example. For the graphs considered here, we found our posterior to converge immediately as soon as initial data $\mathcal{D}$ is provided, or after one/two interventions. CEO can take advantage of this, and simply optimize the MI for $y^\star$.

Healthcare The SCM for this real-world application results in causal effects that are linear functions of their inputs, observed with standard white Gaussian noise. We designed four incorrect graphs which represent plausible hypotheses a doctor may have in this context Appendix J. Due to the simplicity of the underlying true functions, one can see that the greediness of EI (used by CBO) grants it better performance on average. CEO still consistently outperforms CBO on the wrong graphs. This also shows evidence that while exact knowledge of the graph is not always required for efficient optimization, it can be better on average than incorrect knowledge of it.

Epidemiology In this example, adapted from (Havercroft and Didelez, 2012), the goal is to subministrate doses for two potential treatments, which we denote as T and R, (see (Havercroft and Didelez, 2012) for details) to minimize HIV viral load. The associated DAG is shown in Fig. 3. Fig. 4(c) shows how, in this more challenging scenario, both competitor methods perform worse. Indeed, the multimodal nature of the causal effects and the high observation noise characterizing this example penalise CBO and its EI acquisition function. While this has been observed in a-causal BO, it is even more problematic when exploring and comparing multiple functions as in CBO.

Extended Epidemiology: To conclude, we test the algorithm on a larger graph obtained by adding confounders to the Epidemiology DAG. The resulting DAG includes 10 nodes, and the associated wrong graphs are created by removing the same edges from the original graph as in the Epidemiology experiment. Even under this situation, as in the previous Epidemiology experiment, we find that CEO generally performs better than CBO (with both wrong and true). Indeed, a larger graph implies a larger product in the truncated factorization term; this can imply further overconfidence in the CBO surrogate model, whereas CEO averages over multiple plausible hypotheses.

6 DISCUSSION & CONCLUSIONS

We proposed CEO, a framework that allows an experimenter to efficiently solve the causal global optimization problem when the graph is unknown. CEO handles continuous variables with flexible nonparametric relationships, while still allowing for a closed-form posterior over graphs that combines observational and interventional data. Our experiments with synthetic and real-world DAGs show that CEO allows the experimenter to reach the global optimum with significantly reduced cost compared to CBO when the graph is unknown, and sometimes even when the graph is known. Our acquisition, CES, allows to learn the graph only to the extent to which it is useful for optimization, and it is robust to observation noise.

Limitations. A limitation of CEO is its restriction to continuous variables, inherited from CBO. Further, a significant computational effort is required to efficiently approximate the CES objective. In real examples this computational cost may be accepted because the interventions represent a costly experiment (such as administering a drug to a patient).
The formulation of CES can be seen as an active-learning objective of independent interest, and thus opens promising avenues for future work (e.g. joint causal effect estimation and structure learning). Further, our surrogate model definition could also allow for approximating expectations over the graph (e.g. exploiting the rich literature on MCMC with dynamic programming for structure learning (Eaton

and Murphy, 2007a; Zemplenyi and Miller, 2021)) to help scalability of posterior inference in very large graphs. Some form of parameter sharing between surrogate models corresponding to *nested* sets of variables could be used to reduce cost, as suggested by a reviewer.

Acknowledgements

NB acknowledges support from the Alan Turing Institute via the Enrichment Scheme. TD acknowledges support from a UKRI Turing AI Acceleration Fellowship [EP/V02678X/1].

References

R. J. Adler, J. E. Taylor, et al. *Random fields and geometry*, volume 80. Springer, 2007.

V. Aglietti, T. Damoulas, M. Álvarez, and J. González. Multi-task causal learning with gaussian processes. In *Advances in Neural Information Processing Systems*, volume 33, 2020a.

V. Aglietti, X. Lu, A. Paleyes, and J. González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020b.

V. Aglietti, N. Dhir, J. González, and T. Damoulas. Dynamic causal bayesian optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

R. Agrawal, C. Squires, K. Yang, K. Shanmugam, and C. Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 3400–3409. PMLR, 16–18 Apr 2019.

S. A. Andersson, D. Madigan, M. D. Perlman, et al. A characterization of markov equivalence classes for acyclic digraphs. *Annals of statistics*, 25(2):505–541, 1997.

J. Angrist and G. Imbens. Identification and estimation of local average treatment effects, 1995.

O. Atan, J. Jordon, and M. van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

D. M. Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

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

G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 116–125, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606149.

C. C. Drovandi, C. Holmes, J. M. McGree, K. Mengersen, S. Richardson, and E. G. Ryan. Principles of experimental design for big data analysis. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 32(3): 385, 2017.

D. Eaton and K. Murphy. Bayesian structure learning using dynamic programming and mcmc. *Uncertainty in Artificial Intelligence*, 2007a.

D. Eaton and K. Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pages 107–114. PMLR, 2007b.

G. R. A. Faria, A. Martins, and M. A. Figueiredo. Differentiable causal discovery under latent interventions. In *Conference on Causal Learning and Reasoning*, pages 253–274. PMLR, 2022.

P. I. Frazier. Bayesian optimization. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, pages 255–278. INFORMS, 2018.

N. Friedman and D. Koller. Being Bayesian about network structure; a Bayesian approach to structure discovery in Bayesian networks. *Machine learning*, 50(1):95–125, 2003.

N. Friedman and I. Nachman. Gaussian process networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, page 211–219, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607099.

J. L. Gamella and C. Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *arXiv preprint arXiv:2006.05690*, 2020.

R. Garnett. *Bayesian Optimization*. Cambridge University Press, 2022. in preparation.

A. Ghassami, S. Salehkaleybar, N. Kiyavash, and E. Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pages 1724–1733. PMLR, 2018.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

S. Håkansson, V. Lindblom, O. Gottesman, and F. D. Johansson. Learning to search efficiently for causally near-optimal treatments. *arXiv preprint arXiv:2007.00973*, 2020.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4): 926–939, 2014.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):291–318, 2015.

W. Havercroft and V. Didelez. Simulating from marginal structural models with time-dependent confounding. *Statistics in medicine*, 31(30):4190–4206, 2012.

P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6), 2012.

P. Hennig, M. A. Osborne, and H. P. Kersting. *Probabilistic Numerics*. Cambridge University Press, 2022.

S. Horii. Bayesian model averaging for causality estimation and its approximation based on gaussian scale mixture distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 955–963. PMLR, 2021.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Do-calculus when the true graph is unknown. In *UAI*, pages 395–404. Citeseer, 2015.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

M. Kocaoglu, A. Dimakis, and S. Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884. PMLR, 2017a.

M. Kocaoglu, A. G. Dimakis, S. Vishwanath, and B. Hassibi. Entropic causal inference. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017b.

J. Kuipers and G. Moffa. Partition mcmc for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112(517):282–299, 2017.

F. Lattimore, T. Lattimore, and M. D. Reid. Causal bandits: Learning good interventions via causal inference. *arXiv preprint arXiv:1606.03203*, 2016.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Y. Lu, A. Meisami, and A. Tewari. Causal bandits with unknown graph structure. *Advances in Neural Information Processing Systems*, 34, 2021.

D. J. MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995.

A. Marx, L. Yang, and M. van Leeuwen. Estimating conditional mutual information for discrete-continuous mixtures using multi-dimensional adaptive histograms. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 387–395. SIAM, 2021.

A. R. Masegosa and S. Moral. An interactive approach for bayesian network learning using domain/expert knowledge. *International Journal of Approximate Reasoning*, 54(8):1168–1181, 2013.

N. Miklin, A. A. Abbott, C. Branciard, R. Chaves, and C. Budroni. The entropic approach to causal correlations. *New Journal of Physics*, 19(11):113041, 2017.

H. Moss. *General-purpose Information-theoretical Bayesian Optimisation: A thesis by acronyms*. PhD thesis, Lancaster University, 2021.

K. P. Murphy. Active learning of causal bayes net structure. Technical report, 2001.

R. O. Ness, K. Sachs, P. Mallick, and O. Vitek. A bayesian active learning experimental design for inferring signaling networks. In *International Conference on Research in Computational Molecular Biology*, pages 134–156. Springer, 2017.

A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

J. Pearl. *Causality*. Cambridge university press, 2009.

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

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

E. Sgouritsa. *Causal Discovery Beyond Conditional Independencies*. PhD thesis, University of Tubingen, 2015.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

R. Silva. Observational-interventional priors for dose-response learning. *arXiv preprint arXiv:1605.01573*, 2016.

R. Silva and R. B. Gramacy. Gaussian process structural equation models with latent variables. *arXiv preprint arXiv:1002.4802*, 2010.

P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. Aimale, and F. Wimberly. Constructing bayesian network models of gene expression networks from microarray data. 2000a.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000b.

K. Swersky, J. Snoek, and R. P. Adams. Multi-task bayesian optimization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

P. Tigas, Y. Annadani, A. Jesson, B. Schölkopf, Y. Gal, and S. Bauer. Interventions, where and how? experimental design for causal models at scale. *arXiv preprint arXiv:2203.02016*, 2022.

S. Tong and D. Koller. Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence*, volume 17, pages 863–869. Citeseer, 2001.

C. Toth, L. Lorch, C. Knoll, A. Krause, F. Pernkopf, R. Peharz, and J. von Kügelgen. Active bayesian causal inference. *arXiv preprint arXiv:2206.02063*, 2022.

T. Verma and J. Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

J. Viinikka and M. Koivisto. Layering-mcmc for structure learning in bayesian networks. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 839–848. PMLR, 03–06 Aug 2020.

J. von Kügelgen, P. K. Rubenstein, B. Schölkopf, and A. Weller. Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks. In *NeurIPS 2019 Workshop Do the right thing: machine learning and causal inference for improved decision making*. NeurIPS, Dec. 2019.

M. J. Vowels, N. C. Camgoz, and R. Bowden. D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*, 2021.

T.-Z. Wang and Z.-H. Zhou. Actively identifying causal effects with latent variables given only response variable observable. *Advances in Neural Information Processing Systems*, 34, 2021.

Y. Wang, L. Solus, K. D. Yang, and C. Uhler. Permutation-based causal inference algorithms with interventions. *arXiv preprint arXiv:1705.10220*, 2017.

Z. Wang and S. Jegelka. Max-value entropy search for efficient bayesian optimization. In *International Conference on Machine Learning*, pages 3627–3635. PMLR, 2017.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

J. Wilson, F. Hutter, and M. Deisenroth. Maximizing acquisition functions for bayesian optimization. *Advances in neural information processing systems*, 31, 2018.

K. Yang, A. Katcoff, and C. Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

M. Zemplenyi and J. W. Miller. Bayesian optimal experimental design for inferring causal structure. *arXiv preprint arXiv:2103.15229*, 2021.

B. Zhang, A. A. Tsiatis, E. B. Laber, and M. Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

J. Zhang. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pages 11012–11022. PMLR, 2020.

K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7 (2):1–22, 2015.

A NOMENCLATURE

Symbol	Description
$\mathbf{V}$	Observed endogenous variables
$\mathbf{U}$	Set of exogenous background variables
F	Set of deterministic functions
$\mathbf{C}$	Non-manipulative variables
$\mathbf{X}$	Manipulative variables
Y	Output variable
$\mathcal{P}(\mathbf{X})$	Set of all possible interventions which can be performed in the graph
$\mathbf{Pa}_j^{\mathcal{G}}$	Parents of each variable $V_j \in \mathbf{V}$ given by $\mathcal{G}$
N	Number of samples collected from each interventional distribution
pa_i	Denotes the parents of V_i
$\mathbf{X}_I$	One possible intervention set out of a total $\lvert\mathcal{P}(\mathbf{X})\rvert$
$\mathbf{x}_I$	Corresponding values of intervened set of variables $\mathbf{X}_I$
$\mathbf{V}_I$	Corresponding set of non-intervened variables following intervention $\mathbf{X}_I$
$\mathbf{V}_Y$	Denotes $\mathbf{V}_I \setminus Y$
$D(\mathbf{X}_I)$	Intervention domain
ES	Exploration set, equivalent to or subset of $\mathcal{P}(\mathbf{X})$
$\mathcal{D}$	Samples of manifestations of variables in $\mathcal{G}$ following intervention $\mathbf{X}_I$
$\mathbf{X}_I^\star$	Optimal intervention set
$\mathbf{x}_I^\star$	Optimal intervention level(s)
$\mathcal{G}$	True causal graph
G	Graph latent random variable
$P(G)$	Discrete prior over causal graphs
R_G	Support of graph distribution
g	One of the possible graphs in $P(G)$
$m_I(\mathbf{x}_I)$	Mean function used in GP prior on $\mathbb{E}[Y \mid \mathrm{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}]$
$\mathbb{I}(\mathbf{A};\mathbf{B}\mid\mathbf{C})$	Conditional mutual information between sets $\mathbf{A}$ and $\mathbf{B}$ on $\mathbf{C}$
$\mathbb{H}(p(x))$	Entropy if x is discrete, differential entropy if x is continuous
$\mathrm{CO}(\mathbf{X}_I, \mathbf{x}_I)$	Cost of performing intervention $\mathrm{do}(\mathbf{X}_I = \mathbf{x}_I)$
$\mathcal{N}_x(0,1)\Big\rvert_{X=x}$	Evaluate expression (e.g. $\mathcal{N}_x(0,1)$) at $X = x$

B FURTHER DISCUSSION ON GRAPH POSTERIOR AND LIKELIHOOD

Eqs. (14) and (15) can also be seen as *estimators* of the true average causal effect $\mathbb{E}[Y \mid \mathrm{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}]$ and the variance of the interventional distribution $\mathbb{V}[Y \mid \mathrm{do}(\mathbf{X}_I = \mathbf{x}_I), \mathcal{G}]$. Unsurprisingly, the mean estimator minimizes the error for a risk function with MSE loss in this context (Horii, 2021). Both estimators are consistent, in the sense that: (1) as $P(G) \to \delta_{G=\mathcal{G}}$ and as $\widehat{p}(Y \mid \mathrm{do}(\mathbf{X}_I = \mathbf{x}_I), G = g) \to p(Y \mid \mathrm{do}(\mathbf{X}_I = \mathbf{x}_I), G = g)$ (ensured by do-calculus, when the effect is identifiable). Further, as $P(G) \to \delta_{G=\mathcal{G}}$ (a Dirac mass at $\mathcal{G}$), the first term in **??** converges to $\mathbb{V}_{\hat{p}(Y\mid\mathrm{do}(\mathbf{X}_I=\mathbf{x}_I),\mathcal{G})}[Y]$, while the second term vanishes.

Posterior convergence There are standard conditions in our setting that ensure that the graph is identified. The following assumptions in our setting:

1. Causal sufficiency (i.e. no hidden confounders)

2. All variables (except the target) can be manipulated (i.e. intervened on)
3. Infinite samples (observational and interventional) can be obtained from each node
4. Causal minimality; i.e. the joint in line 74 does not Markov-factorize w.r.t. to any sub-graph of $\mathcal{G}$
5. The support of $P(G)$ includes $\mathcal{G}$

are sufficient to guarantee convergence of the posterior to a Dirac mass on the true graph, in the limit as interventions are performed. This is independent of how interventions are collected. That we can intervene on all nodes also guarantees us identifiability of the interventional Markov-equivalence classes (Hauser and Bühlmann, 2015). Finally, we are not aware of finite sample guarantees that apply to our specific setting.

Graph likelihood full expression Denote by $\mathbf{X}^O$ and $\mathbf{y}^O$ the observational inputs and outputs in $\mathcal{D}^O$ and by $pa^g_{j,I}$ the values for the parents of V_j in $G = g$ resulting from an intervention on $\mathbf{X}_I = \mathbf{x}_I^{(i)}$. We can write the likelihood evaluated at a single interventional point $(\mathbf{x}_I^{(i)}, \mathbf{v}_I^{(i)})$ as:

$$\prod_{V_j \in \mathbf{V}_I} \mathcal{N}_{V_j}(m_j, \Sigma_j)\Big|_{V_j = v_j^{(i)}} \quad \text{with} \quad A_j = k_j(pa^g_{j,I}, \mathbf{X}^O)[k_j(\mathbf{X}^O, \mathbf{X}^O) + \sigma_j^2 I]^{-1} \tag{10}$$

$$m_j = A_j \mathbf{y}^O, \quad \Sigma_j = k_j(pa^g_{j,I}, pa^g_{j,I}) - A_j k_j(\mathbf{X}^O, pa^g_{j,I}) \tag{11}$$

where k_j represent the prior kernel functions of f_j and m_j, Σ_j correspond to the standard posterior predictive GP parameters (Williams and Rasmussen, 2006, p.17). Note that, when the intervened variables are parents of V_j, the values of $pa^g_{j,I}$ are replaced by the $\mathbf{x}_I^{(i)}$ values.

Exploring the full space of DAGs Methodologies for exploring the full space of DAGs are well-studied (i.e. defining a posterior over a very large space), and return an approximate solution employing sophisticated MCMC schemes (e.g. over the space of node *orderings*) in the context of Bayesian structure learning (Madigan et al., 1995; Friedman and Koller, 2003; Kuipers and Moffa, 2017; Viinikka and Koivisto, 2020). Our setting is complementary to this research: once a sophisticated approximate posterior is provided via representative samples, this can be incorporated in our framework without adjustments. Therefore, on the structure learning side our work is more related to (von Kügelgen et al., 2019), where also a small number of DAGs, continuous $\mathbf{X}$ and flexible nonparametric priors for SEMs are considered; however, they are not interested in causal optimization.

C CAUSAL ENTROPY SEARCH ACQUISITION COMPUTATION

C.1 Joint entropy details

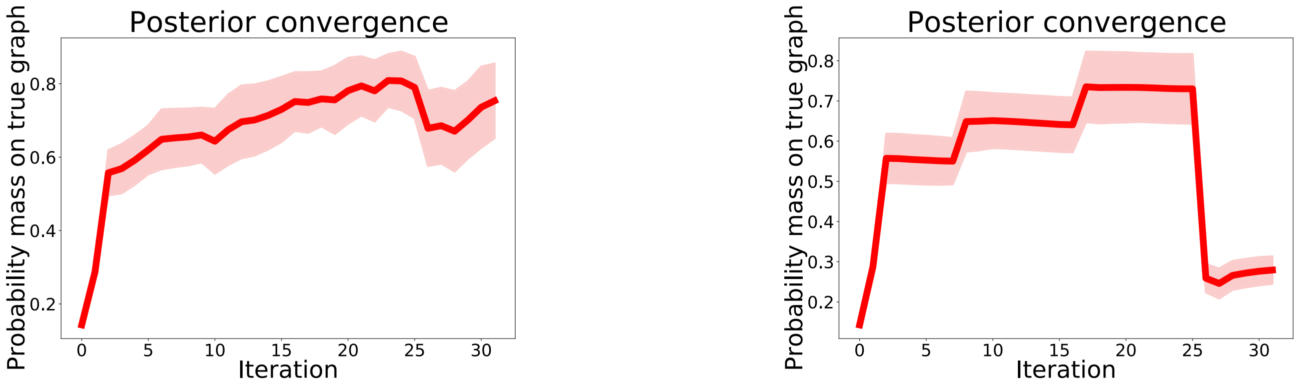
For simplicity we omit conditioning on $\mathcal{D}$ on all terms, and write the joint entropy as:

$$\mathbb{H}[p(y^\star, G)] = -\sum_G \int \mathrm{d}y^\star p(G, y^\star) \log p(G, y^\star) \tag{12}$$

$$= -\sum_G \int \mathrm{d}y^\star p(G|y^\star) p(y^\star) \log p(G|y^\star) - \sum_G \int p(G|y^\star) p(y^\star) \log p(y^\star) \tag{13}$$

$$= -\sum_G \int \mathrm{d}y^\star p(G|y^\star) p(y^\star) \log p(G|y^\star) - \int p(y^\star) \log p(y^\star) \tag{14}$$

We approximate the challenging term (the first) with Monte Carlo samples from our mixture distribution $p(y^\star|\mathcal{D})$ (Eq. (9)) and keep track of the $\mathbf{x}^\star$ associated with each sample of $y^\star$ to update the graph posterior. More accurate approximations could be considered in future work. The second term we approximate as described in Algorithm 2, which details the complete algorithm for CES.



(a)

(b)

Figure 5: Evolution of $P(G = \mathcal{G})$ across iterations in the experiment using the synthetic graph. (a) is the posterior as given by CD-CBO, whereas (b) is that given by CEO. As discussed in §5, we found this an interesting example where the true graph is hard to distinguish from an alternative one, but causal optimization can still proceed efficiently. Note that in both healthcare and epidemiology examples we found the graph to be identified easily by our posterior.

D GRAPH POSTERIOR EVOLUTION PLOTS

E FURTHER DISCUSSION ON GRAPH POSTERIOR AND LIKELIHOOD

Compared to acquisition functions in prior related works (Aglietti et al., 2020b; Wang and Jegelka, 2017; Swersky et al., 2013), several important challenges arise specific to our goals as specified by Eq. (1): to start with, the acquisition used in CBO cannot be straightforwardly extended to our setting. Firstly, because it is not clear how to incorporate our graph prior in a principled manner. Secondly, as explained in Section 3.1, we cannot assume that we can observe the causal effect exactly, since observations from the SEM are noisy, and the family of expected-improvement (EI) acquisitions are known to be inappropriate in this setting (Frazier, 2018; Garnett, 2022). We addressed these challenges by introducing an information-theoretic acquisition function. These types of objectives are widely used in BOED (Drovandi et al., 2017), active learning (MacKay, 1992) as well as many BO approaches (Hennig and Schuler, 2012; Wang and Jegelka, 2017; Moss, 2021). The information-theoretic approach is often advocated in BO because, contrasted to approaches like EI that largely judge optimization performance based *solely* on having found high objective function values, it seeks data that is maximally informative about a variable of interest (Garnett, 2022). In practice, these approaches can be better suited to challenging noisy problems, multimodal and non-smooth functions, or optimization with multiple information sources (Frazier, 2018). It is worth keeping in mind that there is no uniformly best acquisition across all settings.

Output space vs input space ES We decided to frame causal optimization as learning about $y^\star$ rather than $\mathbf{x}^\star$. The question of whether to perform output-space ES (what we do) versus input-space ES has been studied before in BO; see (Garnett, 2022, §6) and (Moss, 2021). However, a crucial difficulty is added in causal optimization: there are *multiple* surrogate models, each with different input space and dimensionality; one for each intervention set $\mathbf{X}_I \in \mathbf{ES}$ we want to consider. Having multiple surrogates is similar to the setting of multi-task BO (Swersky et al., 2013); however (1) our tasks do not share input space and (2) we do not know which task actually contains the global optimum. Therefore, while in principle learning about $\mathbf{x}^\star$ could be done, it would be computationally expensive and inference-wise challenging to define a distribution over a variable with varying (and potentially large) dimensionality. On the other hand, all causal effects share Y, which is one dimensional. Future work could consider inference techniques for dimension-varying parameters like reversible-jump MCMC (Green, 1995).

F FURTHER DISCUSSION ON SURROGATE MODELS

A surrogate model for each graph? Notice that one can think of a very different way to incorporate uncertainty: we could define distinct surrogate models for each g in the support R_G of $P(G)$. We do not take this approach for several reasons: (1) it would not extend to a setting where we need to approximate expectations under $P(G)$ with sampling (hence it is not scalable in this sense); (2) it would require $|R_G| \cdot |\mathbf{ES}|$ kernel hyperparameters to store and update at every iteration; (3) finally, modelling the true causal effect by marginalizing the graph in this approach would lead to a weighted mixture of

GPs, rendering inference and BO complicated.

G BELIEF OVER THE OPTIMAL SET

We compute $P(\mathbf{X}_I = \mathbf{X}_I^\star)$ as:

$$P(\mathbf{X}_I = \mathbf{X}_I^\star) = \frac{e^{\mu_I^\star + \beta \cdot \sigma_I^\star}}{\sum_{I'} e^{\mu_{I'}^\star + \beta \cdot \sigma_{I'}^\star}} \tag{15}$$

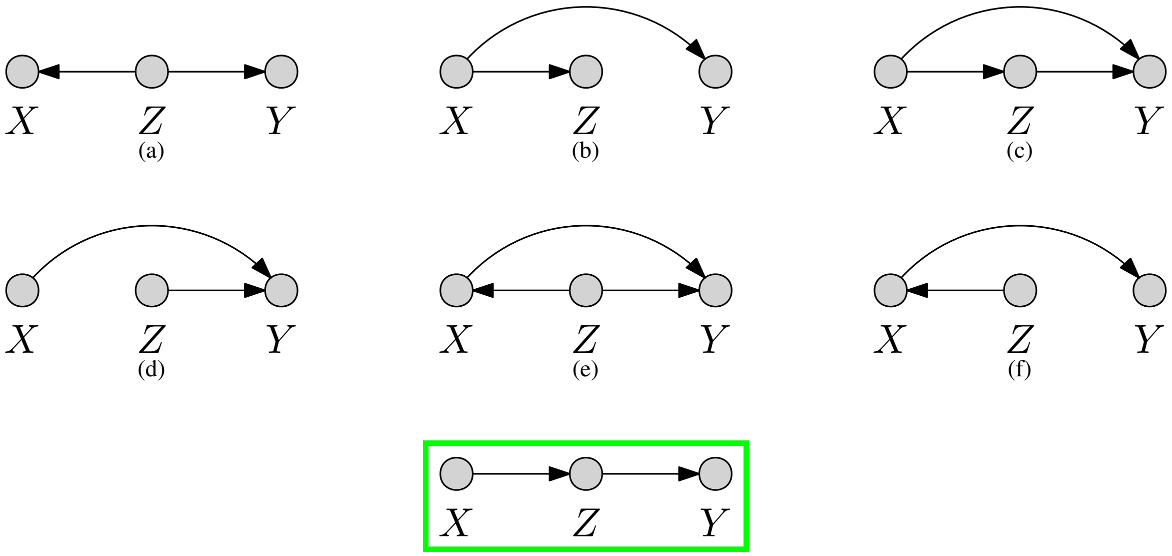
when maximizing a causal effect, with signs reversed when minimizing. Here, $\mu_I^\star$ is the minimum of the mean function of surrogate model f_I, and $\sigma_I^\star$ the standard deviation of the univariate Gaussian with mean at that point. We did not explore adaptive tuning of β which we fixed to 0.1. Guarantees for UCB can be found in e.g. (Garnett, 2022).

G.1 Parametric assumptions

Firstly, the additive Gaussian noise assumption (let us call it AGN- additive Gaussian noise) combined with GPs on the structural equations allows us to define the graph likelihood in closed form. This also implies that GP posteriors on the structural functions are available in closed form. When AGN does not hold, one would need to resort to the literature on Variational GPs, which allows one to perform inference with GPs with non-Gaussian likelihoods. Secondly, the AGN assumption also allows us to deal with noisy Bayesian optimization (Garnett, 2022). Recall that, even if we knew the graph, we do not get to observe $\mathbb{E}[Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), G]$, but can only get samples from $p(Y \mid \text{do}(\mathbf{X}_I = \mathbf{x}_I), G)$. BO with non-Gaussian noise is an active area of research, see e.g. [16] who allow for sub-Gaussian likelihood noises. The optimization becomes more sophisticated, and note that CEO (and CBO) needs to deal with multiple surrogate models (one for each $\mathbf{X}_I$). Therefore, extending these methods from the BO setting to CBO and CEO is a future area of research.
Finally, beyond the previously discussed additive Gaussian noise assumption, in our experiments we used radial basis function kernels, consistently with previous works on CBO. These were appropriate kernels for the SCMs we studied in the experiments; however, for nonstationary causal effect functions one could use nonstationary GP kernels.

H SYNTHETIC: GRAPHS AND TRUE SEM

For the SEM of the synthetic graph, see the Supplementary material of (Aglietti et al., 2020b). For all experiments, we used standard radial basis function (RBF) kernels, and optimized hyperparameters with type II maximum likelihood.



(a) (b) (c) (d) (e) (f) (g)

Figure 6: Graph prior $P(G)$ used for the synthetic experiment s.t. $|R_G| = 6$. The true DAG is shown in Fig. 6g.

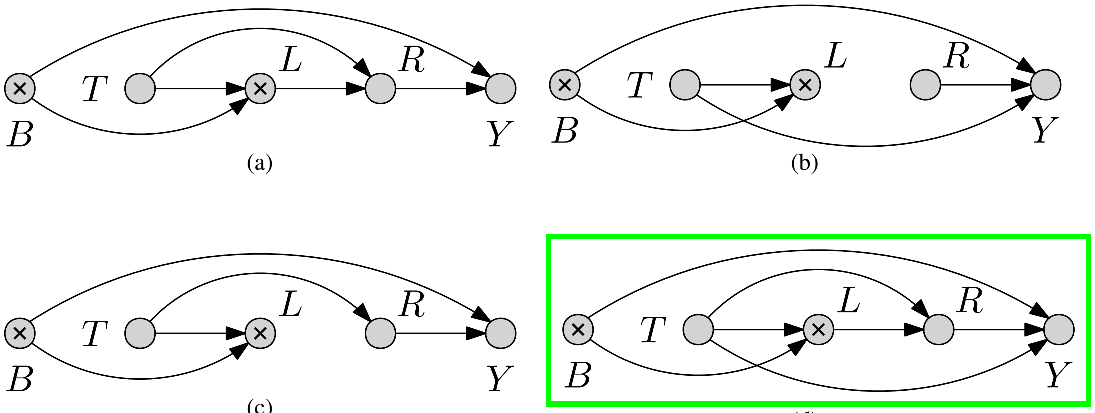
Algorithm 2 CES

```
1:  Inputs:
      • Initial data $\mathcal{D} = (\mathcal{D}^O, \mathcal{D}^I)$
      • Surrogate models: $f_I(\mathbf{x}_I)$ for $\mathbf{X}_I \in \mathbf{ES}$, which have been fitted on $\mathcal{D}$
      • Current graph posterior: $P(G|\mathcal{D})$
      • Intervention sets $\mathbf{ES}$
      • Corresponding acquisition points $\{\mathbf{x}_I^{(s)}\}_{s=1}^S$ (using S acquisition points per set).
2:  Outputs: $\mathbf{X}_I^{\text{best}}, \mathbf{x}_I^{\text{best}}$, maximizers of Eq. (7)
3:  for all $\mathbf{X}_I \in \mathbf{ES}$ do
4:    Get samples $\{y_I^{\star,(j)}\}_{j=1}^J$ approximately distributed from $p(y_I^\star|\mathcal{D})$ using Thompson sampling: $(\mathbf{x}_I^{\star,(j)}, y_I^{\star,(j)}) = \arg\min(\arg\max)_{\mathbf{x}_I} f_I^{(j)}(\mathbf{x}_I), \min(\max)_{\mathbf{x}_I} f_I^{(j)}(\mathbf{x}_I)$ with $f_I^{(j)} \sim \mathcal{GP}(m_I(\mathbf{x}_I), k_I(\mathbf{x}_I, \mathbf{x}_I'))$ as per Eq. (2). Store each $\mathbf{x}_I^{\star,(j)}$, associated to each $y_I^{\star,(j)}$
5:    Fit a KDE estimate $\widehat{p}(y_I^\star|\mathcal{D})$ to the samples $\{y_I^{\star,(j)}\}_{j=1}^J$
6:    Compute $P(\mathbf{X}_I = \mathbf{X}_I^\star)$ as in Appendix G
7:  end for
8:  Obtain samples $\{y_k^\star\}_{k=1}^K$ from $p(y^\star|\mathcal{D})$ by sampling from the mixture in Eq. (9); keeping track of each associated $\mathbf{x}^{\star,k}$
9:  Fit KDE estimate $\widehat{p}(y^\star|\mathcal{D})$ to the samples $\{y^{\star,k}\}_{k=1}^K$
10: Note: up to here, computations do not depend on acquisition points.
11: if $\mathbb{H}[p(G|\mathcal{D})]$ is $\approx 0$ (i.e. CEO is sure that it has found the graph) then
12:   Approximate $\mathbb{H}(p(y^\star|\mathcal{D}))$ with the entropy of the KDE estimate $\mathbb{H}(\widehat{p}(y^\star|\mathcal{D}))$ via quadrature
13:   for all $\mathbf{X}_I \in \mathbf{ES}$ do
14:     for all $\mathbf{x}_I^{(s)} \in \{\mathbf{x}_I^{(s)}\}_{s=1}^S$ do
15:       Update (a copy of) the surrogate model $f_I(\mathbf{x}_I)$ with $\mathbf{x}_I^{(s)}$
16:       Generate fantasy observation by performing $\text{do}(\mathbf{X}_I = \mathbf{x}_I^{(s)})$: sample $\mathbf{v}_Y^l \sim p(\mathbf{v}_Y^l|\text{do}(\mathbf{x}_I^{(s)}))$ using ancestor sampling with SEM functions estimated on $\mathcal{D}$ , whereas $y^l \sim p(Y|\text{do}(\mathbf{x}_I^{(s)}), \mathcal{D})$ is given by the surrogate model on $Y$ , for $l = 1, \dots, L$.
17:       Repeat steps 3 to 9 to get an updated set $\{\widehat{p}(y^\star|\mathcal{D} \cup (\mathbf{x}_I^{(s)}, \mathbf{v}_Y^l, y^l))\}_{l=1}^L$
18:       Compute the average change in entropy, conditioned on fantasy observations: $\frac{1}{L}\sum_{l=1}^L \mathbb{H}(\widehat{p}(y^\star|\mathcal{D})) - \mathbb{H}(\widehat{p}(y^\star|\mathcal{D} \cup (\mathbf{x}_I^{(s)}, \mathbf{v}_Y^l, y^l)))$
19:     end for
20:   end for
21:   Return $\mathbf{x}_I^{(s)}$ maximizing its change in entropy as computed in step 18
22: else if $\mathbb{H}[p(G|\mathcal{D})]$ is not $\approx 0$ then
23:   Compute joint entropies $\mathbb{H}[p(y^\star, G|\mathcal{D})]$ as in Eq. (12) and $y^\star$ samples obtained in step 8
24:   Do steps 13 - 20, but for these joint entropies, for all $\mathbf{x}_I^{(s)}, \mathbf{X}_I$: $\frac{1}{L}\sum_{l=1}^L \mathbb{H}[p(y^\star, G|\mathcal{D})] - \mathbb{H}[p(y^\star, G|\mathcal{D} \cup (\mathbf{x}_I^{(s)}, \mathbf{v}_Y^l, y^l))]$
25: end if
26: Return: $\mathbf{x}^{\text{best}}$ as the one associate with larger entropy reduction among the $\{\mathbf{x}_I^{(s)}\}_{s=1}^S$, and its associated $\mathbf{X}^{\text{best}}$
```

I EPIDEMIOLOGY: GRAPHS AND TRUE SEM

We use a modified (more challenging and nonliner) version of the SEM only partially specified in (Havercroft and Didelez, 2012):

$$B = \mathcal{U}[-1, 1] \tag{16}$$
$$T = \mathcal{U}[4, 8] \tag{17}$$
$$L = \text{expit}(0.5 \cdot T + U) \tag{18}$$
$$R = 4 + L \cdot T \tag{19}$$
$$Y = 0.5 + \cos(4 \cdot T) + \sin(-L + 2 \cdot R) + U + \epsilon \qquad \text{with } \epsilon \ \sim \mathcal{N}(0, 1) \tag{20}$$



(a)

(b)

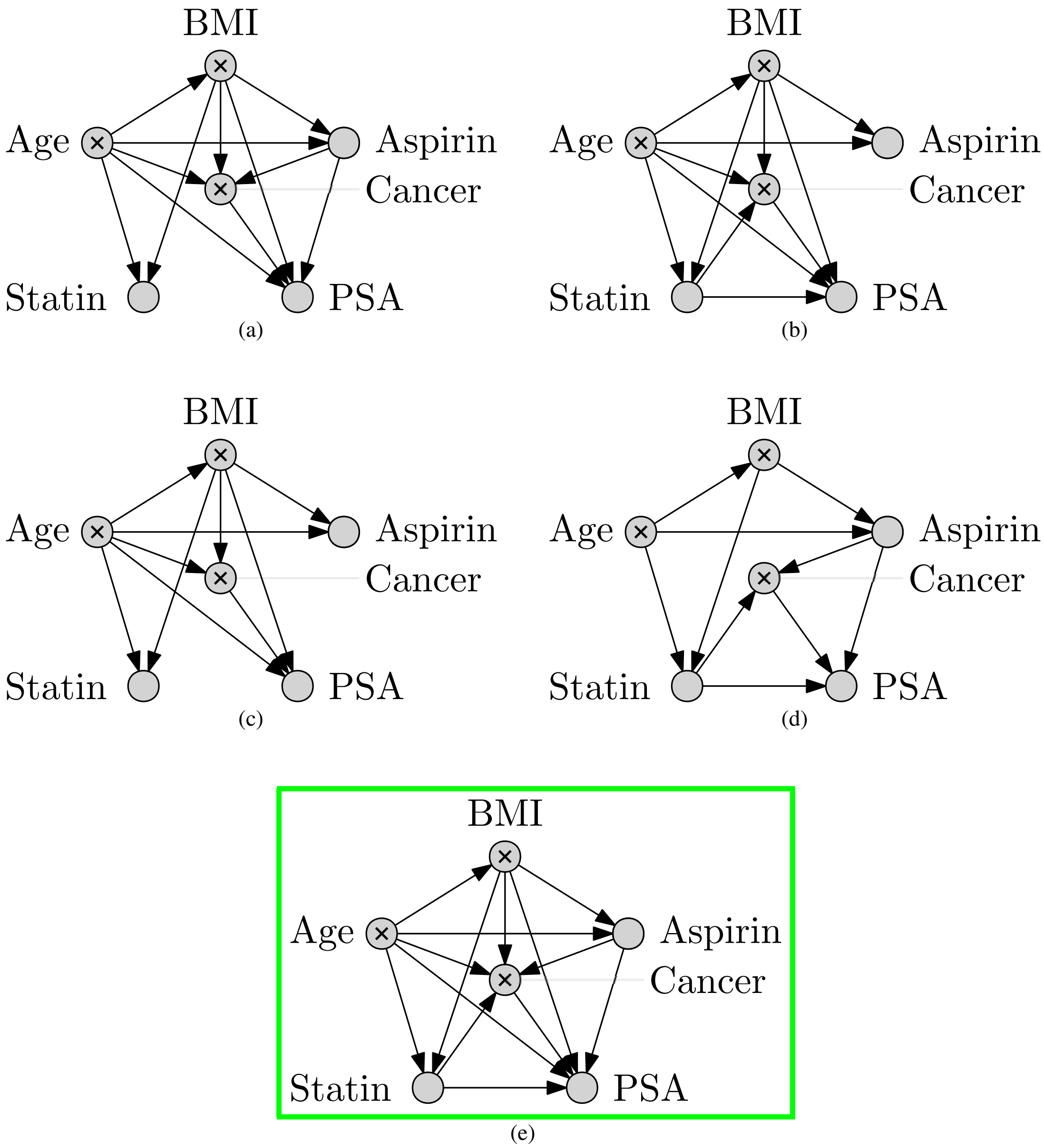
(c)

(d)

Figure 7: Graph prior $P(G)$ used for the epidemiology experiment s.t. $|R_G| = 3$. The true DAG is shown in Fig. 7d.

J HEALTHCARE: GRAPHS AND TRUE SEM

For the SEM see the Supplementary material of (Aglietti et al., 2020b). In this example, we had initial points $|\mathcal{D}^I| = 2$.



(a)

(b)

(c)

(d)

(e)

Figure 8: Graph prior $P(G)$ used for the health experiment s.t. $|R_G| = 4$. The true DAG is shown in Fig. 8e.

K ENTROPY IN CAUSALITY

There is a large literature on the use of information theoretic concepts with applications to causal inference, discovery and counterfactual reasoning. The information-geometric approach to causal discovery (Janzing et al., 2012) attempts to lift the strict conditional independence requirement of classical algorithms like PC (Spirtes et al., 2000b), by defining independence as orthogonality in information space. They find that in important examples this induces a desirable asymmetry between cause and effect. Miklin et al. (2017) find use in entropies between variables of a causal Bayesian network in deriving so-called causal inequalities, which are bounds on certain quantities that characterize the behaviour of the underlying system represented by the CBN. Recently, Kocaoglu et al. (2017b) introduced a framework, "Entropic Causal Inference", where the causal direction between two categorical variable can be discerned from observational data based on an interesting condition on the entropy of the exogenous variable.

Table 3: Average GAP ± one standard error computed across 12 replicates initialized with different $\mathcal{D}$. Higher values are better, and $0 \leq$ GAP ≤ 1. The best result for each experiment in bold.

Method	Synthetic	Epidemiology (Fig. 3)
CEO	**0.75** ± 0.03	0.58 ± 0.05
CBO w. $G = \mathcal{G}$	0.66 ± 0.04	**0.62**± 0.02
CBO w. $G \neq \mathcal{G}$	0.59 ± 0.035	0.46 ± 0.05
CD $-$ CBO	0.39 ± 0.04	0.52 ± 0.04

L COMPUTATIONAL COMPLEXITY

The complexity of CEO is driven by the computation of CES. The parameters influencing these are: $|\mathbf{ES}|$, $\max_I |\mathbf{X}_I|$, the number of acquisition points N (i.e., how many values of the intervened variable to consider, assuming the same number for all $\mathbf{X}_I$). Therefore, the total complexity of the acquisition is $\mathcal{O}(N \cdot \text{CES}(\mathbf{x}) \cdot \sum_{\mathbf{X}_I \in \mathbf{ES}} |\mathbf{X}_I|)$. Here, $\text{CES}(\mathbf{x})$ denotes the time needed to compute CES for a specific value $\mathbf{x}$, regardless of its corresponding $\mathbf{X}_I$. Note this is valid for CBO also, replacing $\text{CES}(\mathbf{x})$ with $\text{CEI}(\mathbf{x})$, the acquisition used by CBO. However,$\text{CEI}(\mathbf{x})$ is cheaper than $\text{CES}(\mathbf{x})$. In practice, CEI operations can be more easily vectorized. CES operations however can be parallelized over both $\mathbf{ES}$ and values of $\mathbf{x}_I$. Our implementation uses KDEs to estimate univariate marginal distributions, while conditionals are estimated with GPs. **Graph sizes**: Since in our setting the number of nodes is not too large, there is a tractable number of graphs that can be enumerated, therefore omitted in the above Big-O notation. As mentioned in Appendix B, larger spaces of graphs could be explored in future work by e.g. sampling with MCMC, as common in the structure learning literature. Work that explores the full space of DAGs with many nodes will necessarily introduce additional approximations. We further discuss scalability in the limitations in Section 6. It is a general limitation currently in causal global optimization problems that one needs to train $|\mathbf{ES}|$ GPs (in general $|\mathbf{ES}|$ will scale as $2^{|\mathbf{X}|}$), which does not scale with large graphs (here, exact GP inference is cubic in the number of collected interventions).

M FURTHER EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

We provide here additional experimental results. In Table 3 and Fig. 9 we show that, as in our previous experiment, CD-CBO performs worse than CEO, and in this case slightly better than CBO on the true graph. Since as we mentioned in §5, in this example we initially found that our acquisition finds the graph too fast (i.e., at initialization), and therefore it would not be possible to compare to CD-CBO, to provide this additional comparison we updated all graph posteriors (of all methods) only with interventions and not with observational data. This only amplifies the signal between the difference among CD-CBO and CBO, and has no other side-effects on the performance comparison; note that if the graph is found immediately, then CD-CBO is CBO.

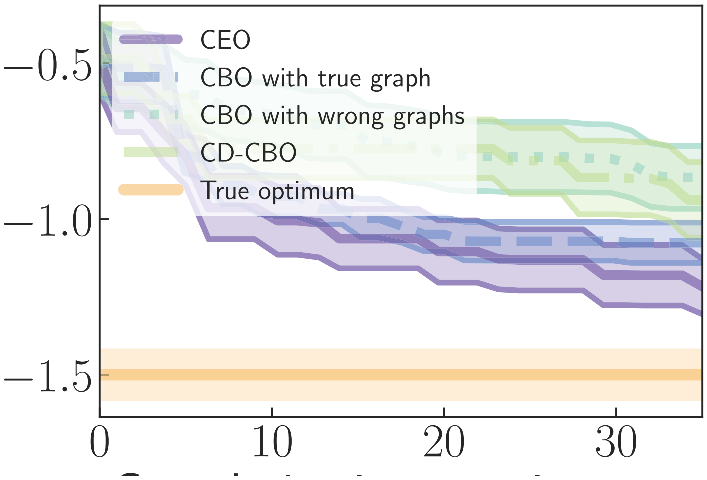


Figure 9: Additional results for a comparision with CD-CBO on the Epidemiology example.