Manuscript version: Author’s Accepted Manuscript
The version presented in WRAP is the author’s accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/173750

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher’s statement:
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.
Causal Entropy Optimization

Nicola Branchini  
University of Edinburgh  
The Alan Turing Institute

Virginia Aglietti  
University of Warwick  
The Alan Turing Institute

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 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 {Statin → Cancer, Statin → PSA, Aspirin → Cancer, Aspirin → PSA}. We propose Causal Entropy Optimization (CEO), a framework which, instead of assuming...
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 $\langle U, V, F, p(U) \rangle$, where $U$ is a set of exogenous background variables distributed according to $p(U)$. $V$ is a set of observed endogenous variables and $F = \{ f_1, \ldots, f_{|V|} \}$ is a set of functions constituting the structural causal model (SCM) such that $v_i = f_i(p_{a_i}, u_i)$ with $p_{a_i}$ denoting the parents of $v_i$. Within $V$, we distinguish between non-manipulable variables $C$ that cannot be intervened on, continuous treatment variables $X$ 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: $p(V \mid \mathcal{G}) = \prod_{V_i \in V} p(V_i \mid Pa_{V_i}^\mathcal{G}, \mathcal{G})$. Denote by $\mathcal{P}(X)$ the power set of $X$ giving all possible interventions we can perform in the graph. The set $X_I \in \mathcal{P}(X)$ represents one such intervention set with $V_I = V \setminus X_I$ 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 $X_I$, the interventional distribution $p(V_I \mid \text{do}(X_I = x_I))$ resulting from setting $X_I$ to a value $x_I$ obeys the following truncated factorization

$$p(V_I \mid \text{do}(X_I = x_I), \mathcal{G}) = \prod_{V_j \in V_I} p(V_j \mid Pa_{V_j}^\mathcal{G})_{|x_I=x_j}$$

where the vertical line represents evaluation of the expression at $X_I = x_I$.

**Notation** Note that, throughout, lowercase denotes realizations of random variables (r.v.), while uppercase denotes the r.v. We denote collected data by $D = \{ \{ (x_i^{(j)}, y_i^{(j)}) \}_{j=1}^N \}_{i=1}^N \in \text{ES}$ where the exploration set (ES) is $\text{ES} = \mathcal{P}(X)$ or $\text{ES} \subset \mathcal{P}(X)$ if only a subset of interventions can be implemented in the system (e.g. if $\mathcal{P}(X)$ is too large). Here, $V_I^{(i)}$ represents one set of values of the non-intervened variables $V_I$ in a mutilated graph where $X_I$ is set to $x_I^{(i)}$. Further, $V_I^{(i)}$ includes both a value for the target variable $y^{(i)}$ and a value for the remaining variables, denoted by $V_I^{(i)} = y^{(i)} \setminus y^{(i)}$. The dataset $D$ includes all data, that is observational data ($D^O$) which correspond to $I = \emptyset$ and interventional data ($D^I$). Every time we intervene in the system we collect $N > 1$ samples from each interventional distribution, but the framework equally handles $N = 1$. See §7 (Appendix) 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 $D$, we seek to identify the interventional set $X_I$ and corresponding values $x_I$ optimizing the causal effect in the true causal graph $\mathcal{G}$

$$x_I^*, x_I^* = \arg\min_{x_I \in \mathcal{P}(X)} \forall x_I \in D(X_I) \mathbb{E}[Y \mid \text{do}(X_I = x_I), \mathcal{G}], \tag{1}$$

where $D(X_I)$ denotes the interventional domain of $X_I$ and the causal effect depends on the true graph $\mathcal{G}$. Solving the problem in Eq. (1) is challenging as evaluating $\mathbb{E}[Y \mid \text{do}(X_I = x_I), \mathcal{G}]$ requires intervening in the real system at a cost, which we assume to be given by $\text{Co}(X_I, x_I)$. 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 $X_I \in ES$ we place a Gaussian process (GP, Williams and Rasmussen, 2006) prior on $g_I(x_I) = \mathbb{E}[Y \mid \text{do}(X_I = x_I), G]$ 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 $g_I(x_I) \sim GP(m_I(x_I), k_I(x_I, x'_I))$, with

$$m_I(x_I) \equiv \mathbb{E}[Y \mid \text{do}(X_I = x_I)],$$

$$k_I(x_I, x'_I) \equiv k(x_I, x'_I) + \mathbb{V}[Y \mid \text{do}(X_I = x_I)],$$

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

$$\mathbb{E}[Y \mid \text{do}(X_I = x_I)] \equiv \mathbb{E}\left[\mathbb{E}[Y \mid \text{do}(X_I = x_I), G]\right]$$

and

$$\mathbb{V}[Y \mid \text{do}(X_I = x_I)] \equiv \mathbb{V}\left[\mathbb{E}[Y \mid \text{do}(X_I = x_I), G]\right] + \mathbb{V}\left[\mathbb{E}[Y \mid \text{do}(X_I = x_I), G]\right],$$

where outer expectations/variances are w.r.t a probability mass function of the r.v. $G$, i.e. $P(G)$, whereas inner expectations/variances are w.r.t the approximation $\mathbb{E}[Y \mid \text{do}(X_I = x_I), G = g]$, a Monte Carlo estimate of an interventional distribution computed via the do-calculus with only observational data.\footnote{This can be computed when the causal effect is identifiable (Pearl, 2009). See §8 (Appendix) for a discussion of the conditions under which Eqs. (2, 3) converge to the true $\mathbb{E}[Y \mid \text{do}(X_I = x_I), G]$ and $\mathbb{V}[Y \mid \text{do}(X_I = x_I), G]$ respectively.}

Surrogate Model Likelihood For each intervention set $X_I \in ES$ and value $x_I^{(i)}$, the output $y$ is a noisy realisation of the objective function $y = g_I(x_I^{(i)}) + \nu$ at $x_I^{(i)}$, where $\nu \sim \mathcal{N}(0, s^2)$. Every time we perform $\text{do}(X_I = x_I^{(i)})$, we obtain a single sample $y^{(i)}$ from the resulting interventional distribution $p(Y \mid \text{do}(X_I = x_I^{(i)}), G)$. 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 $p(g_I \mid D)$ can be derived analytically and will also be a GP with mean $m_I(X \mid D)$ and covariance functions $k_I(x_I, x'_I \mid D)$ computed by standard GP updates (Williams and Rasmussen, 2006).\footnote{This is a common assumption, see e.g. von Kügelgen et al. (2019) or Silva and Gramacy (2010).}

3.2 Inference on the causal graph

Given $D$, we update the prior distribution on $G$ to get $P(G \mid D)$. 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 $P(G = g) = \frac{1}{|G|}$. 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 $P(G \mid D)$ 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 term.\footnote{For every $g$ and $X_j \in X$, we have

$$X_j = f_j(Pa_j^{G=g}) + u_j,$$

where $u_j \sim \mathcal{N}(0, \sigma^2_j)$ and where $\theta_j$ represents a set of hyper-parameters. Further, $x$ and $x'$ are two values for the parents of $V_j$ in $G = g$ and $k_j$ a kernel function of choice. Exploiting the available data we can then compute $p(f_j \mid Pao_j^{G^{D}}, D^{G^O})$ 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 $\mathbb{E}[\cdot]$ (details on likelihood computation in §8.)}

For every $g$ and $X_j \in X$, we have

$$X_j = f_j(Pa_j^{G=g}) + u_j,$$

$$f_j \mid (Pa_j^{G=g} = \theta_j) \sim \text{GP}(0, k_j(x, x'; \theta_j))$$

where $u_j \sim \mathcal{N}(0, \sigma^2_j)$ and where $\theta_j$ represents a set of hyper-parameters. Further, $x$ and $x'$ are two values for the parents of $V_j$ in $G = g$ and $k_j$ a kernel function of choice. Exploiting the available data we can then compute $p(f_j \mid Pa_j^{G}, D^{G^O})$ 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 $\mathbb{E}[\cdot]$ (details on likelihood computation in §8.)

Graph Posterior Given the prior distribution on $G$ and the likelihood, we can compute the posterior distribution on $G$ in closed form as $P(G \mid D \propto p(D \mid G)P(G)$ due to the tractability of GPs. Importantly, this also allows us to compute the expectations w.r.t. $P(G \mid D)$ without approximations in Eqs. (2, 3) and later in the acquisition function.
Convergence Under some conditions, convergence to $\delta_{G^*} = G$ can be guaranteed asymptotically. The key requirements needed are: causal sufficiency; $R_G$ including $G$; all variables can be manipulated; infinite samples can be obtained from each node; causal minimality. See Appendix Section B for details.

Effect of graph knowledge on optimization It is important to emphasize that while computing $P(G \mid D)$ 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. 4 showing different causal graphs and the associated surrogate models, when three interventions are collected. Here the true DAG is given in Fig. 4a (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. 4a 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. 4a are the same.

As demonstrated experimentally in Fig. 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 $P(G \mid D)$ 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 $y^* = \min(\max)_{y^*} y^*_G$ and about $G$. This information gain can be quantified by considering the reduction in uncertainty in an appropriately defined joint distribution on $y^*$ and $G$. While a distribution on the graph, $P(G \mid D)$, was defined in §3.2, the distribution on $y^*$ is implicitly induced by the surrogate models in §3.1.

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

$$X_I^{\text{best}}, x_I^{\text{best}} = \arg \max_{X_I \in \text{ES}, x_I \in D(X_I)} \alpha_{\text{CES}}(X_I, x_I)$$

$$\alpha_{\text{CES}}(X_I, x_I) = \frac{I[y^*, G] : (X_I, x_I, v_Y, y) \mid D]}{\text{Co}(X_I, x_I)}$$

where the denominator $\text{Co}(X_I, x_I)$ 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 $p(y^*, G)$ and the associated posterior. We write its joint entropy
Algorithm 1 CES

1: Input: $D$, $P(G)$, $H$ (N. of iterations)
2: Output: $X_1^*,x_1^*,P(G|D_H)$.
3: Initialise: Set $D_0 = D$.
4: Compute $p(D_0 | G = g)$ for each $g \in R_G$.
5: Compute $P(G | D_0)$.
6: Compute $m_I(x_I)$ and $k_I(x_I, x_I')$ using the do-calculus for each $X_I \in ES$ (Eqs. (4)).
7: for $h = 1, ..., H$ do
8: Compute $\alpha_\text{CES}(X_I, x_I)$ for each $X_I \in ES$ and each $x_I$ in the acquisition points with CES (Algorithm 2 (Appendix)).
9: Obtain the optimal set-value pair $(X_I^h, x_I^h)$.
10: Intervene on the system and augment the dataset $D_h = D_{h-1} \cup (x_I^h, v_I^h)$.
11: Compute $P(G | D_h)$ and every $p(g_I | D_h)$.
12: end for
13: Return $(X_I^*, x_I^*), P(G|D_H)

\[ \mathbb{H}[p(y^*, G | D)] \text{ as } \]
\[ - \sum_G \int dy^* p(G, y^* | D) \log p(G, y^* | D). \] (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
\[ \mathbb{E}[\mathbb{H}[p(y^*, G | D)] - \mathbb{E}[p(y^*, G | D \cup (x_I, v_Y, y))], \]

where the expectation is w.r.t $p(v_Y, y | \text{do}(X_I = x_I), D)$. This is the reduction in entropy observed on average in the joint distribution on $(y^*, G)$ when performing $\text{do}(X_I = x_I)$ and observing $v_Y = (v_Y, y)$. 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 $v_Y$ we would get by simulating what would happen in the system under $\text{do}(X_I = x_I)$. 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 $p(y^*, G)$, and approximate $\mathbb{H}[p(y^*, G)]$.

Joint posterior over $y^*$ and $G$ Computing CES requires defining $p(y^*, G)$, obtaining its posterior given $D$ and computing Eq. (8). We model the joint distribution as $p(y^*, G | D) = P(G | y^*, D)p(y^* | D)$. We can thus write
\[ p(y^* | D) = \sum_{y_I \in ES} P(X_I = X^*) p(y_I^* | D), \]

where $X^*$ is the intervention set in ES which yields the global optimum. We can thus sample from Eq. (9) via standard mixture sampling techniques (Owen, 2013). 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\(^\dagger\). See our Table 1 in the Appendix for connections between the CES and related existing objectives in experimental design and Bayesian optimization.

**Motivation of 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 $X^*$, and $p(y^* | D)$ to turn into $p(y_I^* | D)$ for the $X_I$ s.t. $X_I = X^*$. To model the belief over the optimal set $P(X_I = X^*)$, we employ a *multi-armed bandit* (Lattimore and Szepesvári, 2020) perspective and define the weights via an upper confidence bound (UCB) policy; see §13 (Appendix) for details.

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

\(^\dagger\)Therefore, this formulation provides a useful acquisition that could be used not only for causal optimization, but also in a-causal settings.
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; Horn, 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; Häkansson 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). 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. Note that 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 $D^*$, while $D^0$ is fixed. We set $|D^*_0| = 2$ and $|D^0_0| = 200$ (initial data) unless otherwise specified. 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; $y^*$ denotes the best value of the causal effect across ES.

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.

<table>
<thead>
<tr>
<th>Acquisition</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>CES (ours)</td>
<td>$I[(y^*,G);(X_1,x_1,v_1,y)]</td>
</tr>
<tr>
<td>Entropy Search (Hennig and Schuler, 2012)</td>
<td>$I[(x,y);x^*]</td>
</tr>
<tr>
<td>Max-value Entropy Search (Wang et al., 2017)</td>
<td>$I[(x,y);y^*]</td>
</tr>
<tr>
<td>Multi-Task BO (Swersky et al., 2013)</td>
<td>$I[(x^<em>,y);x^</em>]</td>
</tr>
<tr>
<td>Causal discovery via BO (von Kügelgen et al., 2019)</td>
<td>$I[G;(X_1,x_1,v_1)</td>
</tr>
</tbody>
</table>

* A Python implementation will be released post-review.
We start by considering data generated by the chain graph when learning about the graph. This is to show that jointly CD (2019) for GAP values are better. Variables, are not applicable to the problem setting thus we do not consider them as an alternative to von Kügelgen et al. (2019), and then runs to our continuous variables setting.

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

$$\frac{y(\mathbf{x}^*) - y(\mathbf{x}_{\text{init}})}{y(\mathbf{x}_{\text{best}}) - y(\mathbf{x}_{\text{init}})} + \frac{H - H(\mathbf{x}_{\text{best}})}{H} \left(1 + \frac{H - 1}{H}\right).$$

Here, $y^*$ is the true global optimum, $y(\mathbf{x}_{\text{best}})$ 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, $H(\mathbf{x}^*)$ is the number of iterations to reach the best value achieved. Note that $0 \leq \text{GAP} \leq 1$ and that higher values are better.

### 5.1 Synthetic example

We start by considering data generated by the chain graph $X \rightarrow Z \rightarrow Y$ as per Fig. 2a (green dashed box). We consider all possible alternative DAGs with three nodes as presented in Table 2, where the graph is learned along optimization of the effect, and only to the extent to which it is useful for optimization. In Fig. 2c is very hard to distinguish from $\mathcal{G}$ 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 §19 (Appendix).

### 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), we 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.

### 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 perform better, and 0 ≤ GAP ≤ 1. The best result for each experiment in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Synthetic (Fig. 2a)</th>
<th>Health (Fig. 1)</th>
<th>Epi (Fig. 3)</th>
<th>(Epi + Fig. 9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEO</td>
<td>0.75 ± 0.03</td>
<td>0.62 ± 0.07</td>
<td>0.59 ± 0.06</td>
<td>0.65 ± 0.03</td>
</tr>
<tr>
<td>CBO w. $G = \mathcal{G}$</td>
<td>0.66 ± 0.04</td>
<td>0.71 ± 0.05</td>
<td>0.50 ± 0.05</td>
<td>0.50 ± 0.03</td>
</tr>
<tr>
<td>CBO w. $G \neq \mathcal{G}$</td>
<td>0.59 ± 0.03</td>
<td>0.62 ± 0.06</td>
<td>0.50 ± 0.06</td>
<td>0.53 ± 0.03</td>
</tr>
</tbody>
</table>

this, and simply optimize the MI for $y^*$.

**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). 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, which is known to be more greedy than entropy based one (Frazier, 2018). 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.

A limitation of CEO is its restriction to continuous variables, which inherits from CBO. Further, a significant computational effort is required to efficiently approximate the CES objective. However, 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; Zempleynyi and Miller, 2021)) to help scalability of posterior inference in very large graphs. Future work could attempt to find an optimal or near-optimal experiment design criterion.

## References


Causal Entropy Optimization

Causal Entropy Optimization: Supplementary Material

7 NOMENCLATURE

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>Observed endogenous variables</td>
</tr>
<tr>
<td>U</td>
<td>Set of exogenous background variables</td>
</tr>
<tr>
<td>F</td>
<td>Set of deterministic functions</td>
</tr>
<tr>
<td>C</td>
<td>Non-manipulative variables</td>
</tr>
<tr>
<td>X</td>
<td>Manipulative variables</td>
</tr>
<tr>
<td>Y</td>
<td>Output variable</td>
</tr>
<tr>
<td>(\mathcal{P}(X))</td>
<td>Set of all possible interventions which can be performed in the graph</td>
</tr>
<tr>
<td>(\text{Pa}_j^G)</td>
<td>Parents of each variable (V_j \in V) given by (G)</td>
</tr>
<tr>
<td>(N)</td>
<td>Number of samples collected from each interventional distribution</td>
</tr>
<tr>
<td>(\text{pa}_i)</td>
<td>Denotes the parents of (V_i)</td>
</tr>
<tr>
<td>(X_I)</td>
<td>One possible intervention set out of a total (</td>
</tr>
<tr>
<td>(x_I)</td>
<td>Corresponding values of intervened set of variables (X_I)</td>
</tr>
<tr>
<td>(V_I)</td>
<td>Corresponding set of non-intervened variables following intervention (X_I)</td>
</tr>
<tr>
<td>(V_Y)</td>
<td>Denotes (V_I \setminus Y)</td>
</tr>
<tr>
<td>(D(X_I))</td>
<td>Intervention domain</td>
</tr>
<tr>
<td>(\text{ES})</td>
<td>Exploration set, equivalent to or subset of (\mathcal{P}(X))</td>
</tr>
<tr>
<td>(\mathcal{D})</td>
<td>Samples of manifestations of variables in (G) following intervention (X_I)</td>
</tr>
<tr>
<td>(\mathcal{X}_I^*)</td>
<td>Optimal intervention set</td>
</tr>
<tr>
<td>(x_I^*)</td>
<td>Optimal intervention level(s)</td>
</tr>
<tr>
<td>(G)</td>
<td>True causal graph</td>
</tr>
<tr>
<td>(G)</td>
<td>Graph latent random variable</td>
</tr>
<tr>
<td>(P(G))</td>
<td>Discrete prior over causal graphs</td>
</tr>
<tr>
<td>(R_G)</td>
<td>Support of graph distribution</td>
</tr>
<tr>
<td>(g)</td>
<td>One of the possible graphs in (P(G))</td>
</tr>
<tr>
<td>(m_I(x_I))</td>
<td>Mean function used in (GP) prior on (\mathbb{E}[Y \mid \text{do}(X_I = x_I), G])</td>
</tr>
<tr>
<td>(I(A; B</td>
<td>C))</td>
</tr>
<tr>
<td>(\mathbb{H}(p(x)))</td>
<td>Entropy if (x) is discrete, differential entropy if (x) is continuous</td>
</tr>
<tr>
<td>(\text{Co}(X_I, x_I))</td>
<td>Cost of performing intervention (\text{do}(X_I = x_I))</td>
</tr>
<tr>
<td>(\mathcal{N}_x(0, 1)\big</td>
<td>_{X=x})</td>
</tr>
</tbody>
</table>
8 FURTHER DISCUSSION ON GRAPH POSTERIOR AND LIKELIHOOD

Eqs. (14) and (15) can also be seen as estimators of the true average causal effect $\mathbb{E}[Y \mid \text{do}(X_I = x_I), G]$ and the variance of the interventional distribution $\mathbb{V}[Y \mid \text{do}(X_I = x_I), G]$. Unsurprisingly, the mean estimator minimizes the error for a risk function with MSE loss in this context (Horn 2021). Both estimators are consistent, in the sense that: (1) as $P(G) \rightarrow \delta_{G=G}$ and as $\hat{p}(Y \mid \text{do}(X_I = x_I), G = g) \rightarrow p(Y \mid \text{do}(X_I = x_I), G = g)$ (ensured by do-calculus, when the effect is identifiable). Further, as $P(G) \rightarrow \delta_{G=G}$ (a Dirac mass at $G$), the first term in (?) converges to $\mathbb{V}[\hat{p}(Y \mid \text{do}(X_I = x_I), G)]$, 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 $G$
5. The support of $P(G)$ includes $G$

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 $X^O$ and $y^O$ the observational inputs and outputs in $DO$ and by $pa^g_l$ the values for the parents of $V_j$ in $G=g$ resulting from an intervention on $X_I = x^{(i)}_I$. We can write the likelihood evaluated at a single interventional point $(x^{(i)}_I, y^{(i)}_I)$ as:

$$\prod_{V_j \in \mathcal{V}_I} N_{V_j}(m_j, \Sigma_j) \bigg|_{V_j = y^{(i)}} \text{ with } A_j = k_j(pa^g_l, X^O)[k_j(X^O, X^O) + \sigma^2 I]^{-1}$$  \hspace{1cm} (10)

$$m_j = A_j y^O, \quad \Sigma_j = k_j(pa^g_l, pa^g_l) - A_j k_j(X^O, pa^g_l)$$  \hspace{1cm} (11)

where $k_j$ represent the prior kernel functions of $f_j$ and $m_j, \Sigma_j$ 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 $pa^g_l$ are replaced by the $x^{(i)}_I$ 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). Vinikka 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 $X$ and flexible nonparametric priors for SEMs are considered; however, they are not interested in causal optimization.
9 CAUSAL ENTROPY SEARCH ACQUISITION COMPUTATION

9.1 Joint entropy details

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

$$H[p(y^*, G)] = -\sum_G \int dy^* p(G, y^*) \log p(G, y^*)$$  \hspace{1cm} (12)

$$= -\sum_G \int dy^* p(G|y^*) p(y^*) \log p(G|y^*) - \sum_G \int p(G|y^*) p(y^*) \log p(y^*)$$  \hspace{1cm} (13)

$$= -\sum_G \int dy^* p(G|y^*) p(y^*) \log p(G|y^*) - \int p(y^*) \log p(y^*)$$  \hspace{1cm} (14)

We approximate the challenging term (the first) with Monte Carlo samples from our mixture distribution $p(y^*|\mathcal{D})$ (Eq. (9)) and keep track of the $x^*$ associated with each sample of $y^*$ 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.

10 GRAPH POSTERIOR EVOLUTION PLOTS

Figure 5: Evolution of $P(G = G)$ 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.

11 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 $x^\star$. 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 $X_I \in \mathcal{ES}$ 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 $x^\star$ 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).

12 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 $|R_G| \cdot |\mathcal{ES}|$ 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.

13 BELIEF OVER THE OPTIMAL SET

We compute $P(X_I = X_I^\star)$ as:

$$P(X_I = X_I^\star) = \frac{e^{\mu_I^\star + \beta \cdot \sigma_I^\star}}{\sum_{I'} e^{\mu_{I'}^\star + \beta \cdot \sigma_{I'}^\star}},$$

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 $\beta$ which we fixed to 0.1. Guarantees for UCB can be found in e.g. (Garnett, 2022).

13.1 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 $\mathbb{E}[Y \mid \text{do}(X_I = x_I), G]$, but can only get samples from $p(Y \mid \text{do}(X_I = x_I), G)$. 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 $X_I$). 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.

14 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.
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 \( D = (D^O, D^I) \)
   - Surrogate models: \( f_I(x_I) \) for \( X_I \in ES \), which have been fitted on \( D \)
   - Current graph posterior: \( P(G|D) \)
   - Intervention sets \( ES \)
   - Corresponding acquisition points \( \{x_I^{(s)}\}_{s=1}^S \) (using \( S \) acquisition points per set).

2: Outputs: \( X_I^{\text{best}}, x_I^{\text{best}} \), maximizers of Eq. (7)

3: for all \( X_I \in ES \) do
4:    Get samples \( \{y_I^{*(j)}\}_{j=1}^J \) approximately distributed from \( p(y_I^*|D) \) using Thompson sampling: \( (x_I^{*,(j)}, y_I^{*(j)}) = \arg \min \{\arg \max \}_{x_I} f_I^{(j)}(x_I), \min \{\max \}_{x_I} f_I^{(j)}(x_I) \) with \( f_I^{(j)} \sim GP(m_I(x_I), k_I(x_I, x'_I)) \) as per Eq. (2). Store each \( x_I^{*,(j)} \), associated to each \( y_I^{*(j)} \)
5:    Fit a KDE estimate \( \hat{p}(y_I^*|D) \) to the samples \( \{y_I^{*(j)}\}_{j=1}^J \)
6:    Compute \( P(X_I = X_I^*) \) as in §13
7: end for

8: Obtain samples \( \{y_I^k\}_{k=1}^K \) from \( p(y^*|D) \) by sampling from the mixture in Eq. (9); keeping track of each associated \( x^*,k \)
9: Fit KDE estimate \( \hat{p}(y^*|D) \) to the samples \( \{y^*,k\}_{k=1}^K \)
10: Note: up to here, computations do not depend on acquisition points.
11: if \( \mathbb{H}[p(G|D)] \approx 0 \) (i.e. CEO is sure that it has found the graph) then
12:   Approximate \( \mathbb{H}(p(y^*|D)) \) with the entropy of the KDE estimate \( \mathbb{H}(\hat{p}(y^*|D)) \) via quadrature
13: end for

14: for all \( x_I^{(s)} \in \{x_I^{(s)}\}_{s=1}^S \) do
15:    Update (a copy of) the surrogate model \( f_I(x_I) \) with \( x_I^{(s)} \)
16:    Generate fantasy observation by performing \( do(X_I = x_I^{(s)}) \): sample \( v_I^l \sim p(v_I^l|do(x_I^{(s)})) \) using ancestor sampling with SEM functions estimated on \( D \), whereas \( y^l \sim p(Y|do(x_I^{(s)}), D) \) is given by the surrogate model on \( Y \), for \( l = 1, \ldots, L \).
17:    Repeat steps 3 to 9 to get an updated set \( \{\hat{p}(y^*|D \cup (x_I^{(s)}, v_Y^l, y^l))\}_{l=1}^L \)
18:    Compute the average change in entropy, conditioned on fantasy observations: \( \frac{1}{L} \sum_{l=1}^L \mathbb{H}(\hat{p}(y^*|D)) - \mathbb{H}(\hat{p}(y^*|D \cup (x_I^{(s)}, v_Y^l, y^l))) \)
19: end for
20: end for
21: Return \( x_I^{(s)} \) maximizing its change in entropy as computed in step 18
22: else if \( \mathbb{H}[p(G|D)] \approx 0 \) then
23:   Compute joint entropies \( \mathbb{H}[p(y^*, G|D)] \) as in Eq. (12) and \( y^* \) samples obtained in step 8
24: Do steps 13 - 20, but for these joint entropies, for all \( x_I^{(s)}, X_I: \frac{1}{L} \sum_{l=1}^L \mathbb{H}(p(y^*, G|D)) - \mathbb{H}[p(y^*, G|D \cup (x_I^{(s)}, v_Y^l, y^l))] \)
25: end if
26: Return: \( x^{\text{best}} \) as the one associate with larger entropy reduction among the \( \{x_I^{(s)}\}_{s=1}^S \), and its associated \( X^{\text{best}} \)
15  EPIDEMIOLOGY: GRAPHS AND TRUE SEM

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

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

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

16  HEALTHCARE: GRAPHS AND TRUE SEM

For the SEM see the Supplementary material of [Aglietti et al., 2020b]. In this example, we had initial points \(|D^f| = 2\).
17 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 D. Higher values are better, and 0 ≤ GAP ≤ 1. The best result for each experiment in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Synthetic</th>
<th>Epidemiology (Fig. 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEO</td>
<td>0.75 ±0.03</td>
<td>0.58 ± 0.05</td>
</tr>
<tr>
<td>CBO w. G = G</td>
<td>0.66 ± 0.04</td>
<td>0.62 ± 0.02</td>
</tr>
<tr>
<td>CBO w. G ≠ G</td>
<td>0.59 ± 0.035</td>
<td>0.46 ± 0.05</td>
</tr>
<tr>
<td>CD − CBO</td>
<td>0.39 ± 0.04</td>
<td>0.52 ± 0.04</td>
</tr>
</tbody>
</table>

18 COMPUTATIONAL COMPLEXITY

The complexity of CEO is driven by the computation of CES. The parameters influencing these are: |ES|, max |X|, the number of acquisition points N (i.e., how many values of the intervened variable to consider, assuming the same number for all X). Therefore, the total complexity of the acquisition is \( O(N \cdot CES(x) \cdot \sum_{X \in ES} |X|) \). Here, CES(x) denotes the time needed to compute CES for a specific value x, regardless of its corresponding X. Note this is valid for CBO also, replacing CES(x) with CEI(x), the acquisition used by CBO. However, CEI(x) is cheaper than CES(x). In practice, CEI operations can be more easily vectorized. CES operations however can be parallelized over both ES and values of x. 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 §8, 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 |ES| GPs (in general |ES| will scale as \( 2^{|X|} \)), which does not scale with large graphs (here, exact GP inference is cubic in the number of collected interventions).

19 FURTHER EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

We provide here additional experimental results. In Table 2 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 comparison with CD-CBO on the Epidemiology example.