Causal Elicitation for Bayesian Optimization

Jakob Zeitler[1] **Raul Astudillo**[2]

[1]University College London, UK,
[2]Caltech, US,

Abstract

Causal Inference allows scientists and businesses to draw causal conclusions about e.g. their drug-development or marketing campaign. Causal Entropy Search Branchini et al. [2023] was introduced as a way to learn both the causal graph as well as optimise an intervention of interest at the same time. It combines Bayesian Optimisation with the Causal Inference Framework to identify the right molecule or marketing tagline. Here, we present initial work on a crucial extension of CEO, namely the introduction of preference elicitation, an increasingly popular technique in Bayesian Optimisation to elicit crucial causal knowledge from subject matter experts. We introduce the problem of Causal Elicitation for Bayesian Optimisation, discuss elicitation strategies and initial work on empirical evaluation.

1 INTRODUCTION

The Causal Bayesian Optimisation (CBO) literature had it's start with Aglietti et al. [2020] which evaluated the impact of the knowledge of a causal graph on Bayesian Optimisation (BO) tasks, demonstrating that without causal knowledge, BO performs sub-optimally in some applications. The work was then extended to dynamic settings, as well as settings to learn the graph simultaneously with the intervention optimisation Branchini et al. [2023]. For a comparison of recent CBO work, see Appendix 1.

2 PROBLEM SETTING

Our model is similar to the one pursued by Branchini et al. [2023]. It is comprised of an unknown directed acyclic graph, G, and a tuple $\langle \mathbf{U}, \mathbf{V}, F, p(\mathbf{U}) \rangle$, where $\mathbf{V}$ is a set of observed endogenous variables and $F = \{f_1, \ldots, f_{|V|}\}$ is a set of unknown functions characterizing the structural causal model such that $v_i = f(pa_i, u_i)$, where pa_i denotes the parents of V_i. We also let $\mathbf{X} \subset \mathbf{V}$ denote the set of treatment variables that can be set to any specific value within $D(\mathbf{X}_I)$. Finally, we let $Y \in \mathbf{V}$ denote the outcome of interest to be maximized.

As is typical, we assume G is Markovian so that $p(\mathbf{V} \mid G) = \prod_{V_j \in \mathbf{V}} p(V_j \mid pa_j^G, G)$. We also assume causal sufficiency and perfect interventions so that

$$p(\mathbf{V}_I \mid (\mathbf{X} = \mathbf{x}_I), G) = \prod_{V_j \in V_I} p(V_j \mid pa_j^G, \mathbf{X}_I = \mathbf{x}_I)$$

for any $\mathbf{X}_I \subseteq \mathbf{X}$, where $\mathbf{V}_I = \mathbf{V} \setminus \mathbf{X}_I$. Our goal is to solve

$$\max_{\mathbf{X}_I \in \mathcal{P}(\mathbf{X}), \mathbf{x}_I \in D(\mathbf{X}_I)} [Y \mid (\mathbf{X}_I = \mathbf{x}_I), G]$$

3 FRAMEWORK

3.1 STATSITICAL MODEL OVER OUTCOME OBSERVATIONS

For each $\mathbf{X}_I \in \mathcal{P}(\mathbf{X})$, we place a Gaussian process prior over $h_I(\mathbf{x}_I) = [Y \mid (\mathbf{X}_I = x_I), G]$. Following Branchini et al. [2023], we use the following prior mean and covariance functions $m_I(\mathbf{x}_I) = [[Y \mid (\mathbf{X}_I = \mathbf{x}_I)]]$, and

$$k_I(\mathbf{x}_I, \mathbf{x}'_I) = k(\mathbf{x}_I, \mathbf{x}'_I) + \left[\widehat{V}[Y \mid (\mathbf{X}_I = \mathbf{x}_I)]\right] + V\,[[Y \mid (\mathbf{X}_I = \mathbf{x}_I)]]\,,$$

respectively, where the outer expectation and covariance operators are over the posterior over G given the available observational and interventional data, and the inner expectation and covariance operators are computed with respect to $\widehat{p}(Y \mid (\mathbf{X}_I = \mathbf{x}_I), G)$, an approximation of the interventional distribution computed via the do-calculus with only observational data. We assume that observations are of the form $y = h_I(\mathbf{x}_I) + \nu_I$, where $\nu \sim N(0, s_I^2)$, where ν_I is independent across observations. For each I, the posterior on $h_I(\mathbf{x}_I)$ given G and the outcome observations can be computed in closed form using the standard GP regression formulas.

3.2 STATISTICAL MODEL OVER G

We place a uniform prior over G with support $\mathcal{G} = \{g_1, \ldots, g_K\}$. In our case, we have two sources of information to estimate G: interventional (and observational) data and expert information. We assume these two sources are independent and specify a likelihood for each source.

For the interventional data, we use the model of Branchini et al. [2023]. The conditional distribution of X_j given that $G = g$ is given by $X_j = f_j^g(pa_j^g) + \epsilon_j$, where f_j^g is a Gaussian process and $\epsilon_j \sim N(0, \sigma_j^2)$ is independent across observations. For the expert data, we adopt the Bernoulli model of Ibrahim et al. [2023]. The user response $r_{i,j}$ for the existence of an edge $X_i \to X_j$ (i.e, $X_i \in pa_j$) given that $G = g$ is Bernoulli with mean $\mu I\{X_i \in pa_j^g\} + (1-\mu) I\{X_i \notin pa_j^g\}$. Since the likelihood factorizes across outcome observations and expert information, the posterior over G given these two data sources can be computed in closed form.

3.3 CAUSAL ELICITATION STRATEGY

We begin by introducing additional notation. Let $_n$ denote the data collected after n interventions and $\mathcal{E}_m$ denote the data collected after m interactions with the expert. We denote the expectation given $_n$ and $\mathcal{E}_m$ by $_{n,m}$. To motivate our causal elicitation strategy, we consider the following hypothetical question. If we had to commit to an intervention after collecting data of n interventions and m interactions with the expert, how would we choose it? There is not a single answer to this question, but we argue that a sensible choice is to pick an intervention from the set

$$_{\mathbf{X}_I \in \mathcal{P}(\mathbf{X}), \mathbf{x}_I \in D(\mathbf{X}_I)} {}_{n,m}[f_I(\mathbf{x})]$$

Under this premise,

$$\max_{\mathbf{X}_I \in \mathcal{P}(\mathbf{X}), \mathbf{x}_I \in D(\mathbf{X}_I)} {}_{n,m}[f_I(\mathbf{x}) \mid (X_i, X_j, r_{i,j})] - \max_{\mathbf{X}_I \in \mathcal{P}(\mathbf{X}), \mathbf{x}_I \in D(\mathbf{X}_I)} {}_{n,m}[f_I(\mathbf{x})]$$

can be interpreted as the benefit of asking the expert one additional query (X_i, X_j) and observing the response $r_{i,j}$. Our causal elicitation strategy selects at every iteration the query that maximizes the above quantity in expectation given the information available so far, i.e.,

$$(X_i^*, X_j^*) \in _{(X_i, X_j)} {}_{n,m} [\max_{\mathbf{X}_I \in \mathcal{P}(\mathbf{X}), \mathbf{x}_I \in D(\mathbf{X}_I)} {}_{n,m}[f_I(\mathbf{x}) \mid (X_i, X_j, r_{i,j})] - \max_{\mathbf{X}_I \in \mathcal{P}(\mathbf{X}), \mathbf{x}_I \in D(\mathbf{X}_I)} {}_{n,m}[f_I(\mathbf{x})]$$

4 EXPERIMENTS

As a first step to our implementation, we evaluated current implementations of learning DAGs and optimising interventions, such as Branchini et al. [2023]. Unfortunately, so far, we were not able to replicate results found there, yet, possibly due to a typo in their implementation. [1]

Branchini et al. [2023] in their synthetic study use a example from Aglietti et al. [2020] Figure 3, defined as

$$\begin{aligned} X &= \epsilon_X \\ Z &= \exp(-X) + \epsilon_Z \\ Y &= \cos(Z) - \exp(-\frac{Z}{20}) + \epsilon_Y \end{aligned}$$

With the typo removed, we were not able to reproduce the expected results of CEO, possibly due to several reasons:

- **Numerical instability:** CEO requires a choice of anchor points for the acquisition function, for which we chose the default that were used in the original paper. We can increase those, but so does the runtime which quickly becomes prohibitive.
- **Implementation issues:** It's possible the CEO has itself implementation issues that prevent a proper evaluation.

Due to the two reasons above, we decided to not base our empirical evaluation on the CEO code, but are rewriting it from the ground up. Our goal is then to evaluate the impact of expert knowledge on optimisation performance. For that, we will simulate a problem and introduce expert knowledge ranging from fully correct knowledge to completely wrong knowledge. With that, we will also be able to evaluate the failure mode of our method, i.e. the impact of wrong expert knowledge on convergence.

5 CONCLUSION

As causal expert knowledge is crucial to any causal inference ("no causes in, no causes out"), supplementing current Causal Bayesian Optimisation methods via Preference Elicitation is an essential challenge to overcome for better decision making in science and business. So far, we have stated the problem, strategies for how to include preference elicitation and described challenges in implementation. Going ahead, we are rewriting the current CEO code and will run evaluations on the impact of expert knowledge.

[1] see the '4' line 10 in https://shorturl.at/oqS28

References

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020.

Raj Agrawal, Chandler Squires, Karren Yang, Karthik Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery, 2019.

Nicola Branchini, Virginia Aglietti, Neil Dhir, and Theodoros Damoulas. Causal entropy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 8586–8605. PMLR, 2023.

Juan L. Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability, 2021.

Nazaal Ibrahim, ST John, Zhigao Guo, and Samuel Kaski. Targeted causal elicitation. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2023.

Kevin P. Murphy. Active learning of causal bayes net structure. 2006.

Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C. Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning neural causal models with active interventions, 2022.

Scott Sussex, Andreas Krause, and Caroline Uhler. Near-optimal multi-perturbation experimental design for causal structure learning, 2021.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale, 2022.

Simon Tong and Daphne Koller. Active learning for structure in bayesian networks. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, page 863–869, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608125.

Julius von Kügelgen, Paul K Rubenstein, Bernhard Schölkopf, and Adrian Weller. Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks, 2019.

Causal Elicitation for Bayesian Optimization (Supplementary Material)

Jakob Zeitler[1] **Raul Astudillo**[2]

[1]University College London, UK,
[2]Caltech, US,

A COMPARISON TABLE

Method	Struct.							
learning	Effect opt.	Nonliner	BOED	Scalable	Continuous	Finite Data	Value setting	**Expert**
Murphy [2006], Tong and Koller [2001]	✓			✓			✓	
Agrawal et al. [2019]	✓			✓	✓	✓	✓	
Scherrer et al. [2022]	✓		✓			✓	✓	
Gamella and Heinze-Deml [2021]	✓		✓			✓	✓	
von Kügelgen et al. [2019]	✓		✓	✓		✓	✓	✓
Sussex et al. [2021]	✓			✓		✓		
Tigas et al. [2022]	✓		✓	✓	✓	✓	✓	✓
Ours	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Based on Tigas et al. [2022], extended with column for **expert** knowledge