DISCUSSION ON "DYNAMIC CAUSAL BAYESIAN OPTIMIZATION" BY AGLIETTI ET AL.

Shuo Hao
The Hong Kong Polytechnic University
cee-shuo.hao@connect.polyu.hk

January 8, 2025

ABSTRACT

This document presents a comprehensive analysis of the paper "Dynamic Causal Bayesian Optimization" by Aglietti et al. [1], supplemented by a replication of their synthetic experiment using the TensorFlow Probabilistic framework. Written in response to the requirements of the AI & Data Science Associate Internship Program interview, this report provides a detailed, point-by-point examination of the key concepts, methodologies, and findings of the article, alongside insights gained from reproducing the results of the authors.

1 Introduction

Following the introduction of the Causal Bayesian Optimization method [2], Aglietti et al. subsequently proposed the Dynamic Causal Bayesian Optimization (DCBO) approach [1]. The DCBO approach focuses on finding the sequential determination of interventions to optimize a target variable within a dynamic system, which can be modeled using causality [3]. At each time step t, given causal graph $\mathcal{G}_t$ and structural equation model (SEM) M_t, the optimization problem of solved by DCBO is formulated as:

$$\mathbf{X}_{s,t}^{\star}, \mathbf{x}_{s,t}^{\star} = \underset{\mathbf{X}_{s,t} \in \text{MIS}(\mathbf{X}_t), \mathbf{x}_{s,t} \in D(\mathbf{X}_{s,t})}{\operatorname{argmin}} \mathbb{E}\left[Y_t | \text{do}(\mathbf{X}_{s,t} = \mathbf{x}_{s,t}), \mathbb{1}_{t>0} \cdot I_{0:t-1}\right], \tag{1}$$

where $\mathbf{X}_{s,t}^{\star}$ and $\mathbf{x}_{s,t}^{\star}$ are the optimal intervention set and corresponding intervention level, respectively. $\mathbf{X}_t$ denotes manipulative variables, and Y_t represents the target variable. $\text{MIS}(\mathbf{X}_t)$ refers to all minimal intervention sets for $\mathcal{G}_t$ with respect to Y_t, which can be identified based on the Minimality proposition proposed by Lee and Bareinboim [4]. $D(\mathbf{X}_{s,t})$ defines the intervention domain of $\mathbf{X}_{st}$. $\mathbb{1}_{t>0}$ is an indicator function that equal to one when $t > 0$ and zero otherwise. $I_{0:t-1} = \bigcup_{i=0}^{t-1} \text{do}(\mathbf{X}_{s,t}^{\star} = \mathbf{x}_{s,t}^{\star})$ denotes the previous interventions.

Within the Bayesian optimization framework, addressing the problem in Eq. 1 necessitates the construction of a surrogate model using Gaussian processes for each minimal intervention set. The DCBO methodology employs the concepts of causal Gaussian processes, initially introduced in the work on Causal Bayesian Optimization. Specifically, at each time step, the priors of the causal Gaussian processes, encompassing both mean and covariance functions, are derived from samples generated through an estimated SEM, denoted as $\hat{M}_t$, which is itself constructed based on observational data. By incorporating these informative priors, the surrogate models of $\mathbb{E}\left[Y_t | \text{do}(\mathbf{X}_{s,t} = \mathbf{x}_{s,t}), \mathbb{1}_{t>0} \cdot I_{0:t-1}\right]$ are enhanced in their ability to determine optimal intervention points and effectively mitigate the risk of converging to local minima.

A critical challenge in sequential determination of optimal interventions is how to effectively account for past interventions $I_{0:t-1}$ for $t > 0$. To address this, Aglietti et al. proposed three practical assumptions: (i) invariance of causal structure, (ii) additivity of functional mapping for Y_t in M_t, i.e., f_{Y_t}, and (iii) absence of unobserved confounders, as detailed in Assumption 1 of their paper. Although the necessity and practicality of some of these assumptions have been questioned (see Subsection 3.2), it can be theoretically proven that the intervention function

$\mathbb{E}\left[Y_t|\text{do}(\mathbf{X}_{s,t}=\mathbf{x}_{s,t}), \mathbb{1}_{t>0}\cdot I_{0:t-1}\right]$, denoted as $f_{s,t}: D(\mathbf{X}_{s,t}) \to \mathbb{R}$, can be reformulated as:

$$f_{s,t}(\mathbf{x}) = f_Y^{\text{Y}}(\mathbf{f}^\star) + \mathbb{E}_{p(\mathbf{w}|\text{do}(\mathbf{X}_{s,t}=\mathbf{x}), I_{0:t-1})}\left[f_Y^{\text{NY}}(\mathbf{x}^{\text{PY}}, \mathbf{i}^{\text{PY}}, \mathbf{w})\right], \tag{2}$$

where $\mathbf{f}^\star$ is the set of previously observed optimal targets. $\mathbf{x}^{\text{PY}} \in \mathbf{X}_{s,t} \cap pa(Y_t)$ and $\mathbf{i}^{\text{PY}} \in I_{0:t-1} \cap pa(Y_t)$ denote the intervention targets and the previously intervened variables that are parents of Y_t, respectively. The variable $\mathbf{w}$ represents the parents of Y_t that are neither intervention targets nor previously intervened variables. The conclusion in Eq. 2 represents a theoretical contribution of the DCBO methodology, enabling a novel approach to constructing prior mean and covariance functions for causal Gaussian process models (compared to that in the work of Causal Bayesian Optimization). Practically, based on observational data, one could model f_Y^{Y}, f_Y^{NY}, and the SEM using multiple Gaussian process models, thus evaluating the distribution of $f_{s,t}(\mathbf{x})$, which could then be used as priors for causal Gaussian processes.

For Bayesian optimization practitioners, the incorporation of causality can sometimes be highly significant. When the target variable is influenced by multiple variables with causal relationships, causal Bayesian optimization can reduce the dimensionality of the optimization space, alleviate complexity, and more effectively avoid local minima compared to traditional Bayesian optimization methods. In dynamic optimization problems, where the system dynamics can be modeled using causal relationships and correlations exist across both temporal and event dimensions, the DCBO methodology, particularly the conclusion presented in Eq. 2, offers a critical advancement. Essentially, DCBO transforms the dynamic problem into a series of static Causal Bayesian Optimization problems at each time step. By leveraging observations and the insights from Eq. 2, an informative prior can be established, significantly reducing the complexity and convergence cost at each optimization step.

2 Evaluation of the acquisition function

The acquisition function utilized in the paper is the expected improvement (EI), which is of the form:

$$\text{EI}_{s,t}(\mathbf{x}) = \mathbb{E}_{p(y_{s,t})}\left[\max(y_{s,t} - y_t^\star, 0)\right] / \texttt{cost}(\mathbf{X}_{s,t}, \mathbf{x}_{s,t}). \tag{3}$$

Essentially, EI measures how much the current best value is expected to improve per cost if a new point is to be sampled. In Eq. 3, the improvement at $\mathbf{x}$ is represented by $\max(y_{s,t} - y_t^\star, 0)$. By dividing this improvement by $\texttt{cost}(\mathbf{X}_{s,t}, \mathbf{x}_{s,t})$, the causal acquisition function naturally balances the expected gains with the expense of performing each evaluation.

However, Eq. 3 could be considered *incomplete* because the improvement function $\max(y_{s,t} - y_t^\star, 0)$ is specific to maximization tasks, where improvements occur when candidate point yields a value higher than the current optimal. If the task is minimization, the corresponding improvement function should be $\max(y_t^\star - y_{s,t}, 0)$, reflecting how reductions in the objective value constitute 'improvements'. Hence, any practical formulation of the EI per cost must specify whether the task is to maximize or minimize and adapt the improvement term accordingly.

Implementation of the acquisition function on each candidate requires a conditional statement over the task type to determine the form of the improvement function. For the sake of brevity, the determined improvement function is denoted as $\mathcal{I}$, which is a stochastic functional with respect to the input $\mathbf{x}_{s,t}$. To calculate the expectation of $\mathcal{I}(\mathbf{x}_{s,t})$, that is, $\mathbb{E}_{p(y_{s,t})}[\mathcal{I}(\mathbf{x}_{s,t})]$, we integrate $\mathcal{I}$ into the posterior distribution $p(y_{s,t})$ provided by the causal Gaussian process. This integral has a closed-form solution [5] because $p(y_{s,t})$ is a Gaussian distribution [6], with mean and standard deviation denoted as $m_{s,t}$ and $s_{s,t}$, respectively. The solution is given by:

$$\begin{aligned}
\mathbb{E}_{p(y_{s,t})}[\mathcal{I}(\mathbf{x}_{s,t})] &= (m_{s,t} - y^\star)\Phi(z) + s_{s,t}\phi(z) \quad \text{for maximization tasks,} \\
\mathbb{E}_{p(y_{s,t})}[\mathcal{I}(\mathbf{x}_{s,t})] &= (y^\star - m_{s,t})\Phi(z) + s_{s,t}\phi(z) \quad \text{for minimization tasks,}
\end{aligned} \tag{4}$$

where $z = (m_{s,t} - y^\star)/s_{s,t}$ for maximization and $z = (y^\star - m_{s,t})/s_{s,t}$ for minimization; Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, respectively. Furthermore, for simplicity, the cost function in the acquisition function is assumed to be a constant for any input $(\mathbf{X}_{s,t}, \mathbf{x}_{s,t})$ in our demonstration code. However, in practical applications, it should be determined by taking into account realistic constraints and considerations.

Beyond the form of EI acquisition adopted by Aglietti et al., there are still many acquisition strategies available [5], each exhibiting unique strengths and limitations across diverse problem instances. This diversity underscores the inherent complexity of Bayesian optimization tasks, where factors such as dimensionality, landscape modality, and constraint structures can significantly influence the effectiveness of a given strategy. Empirical observations have consistently shown that no single acquisition method universally outperforms others in every scenario. Hence, as a pivotal component of Bayesian optimization, the form of acquisition function should be selected based on the specific requirements and characteristics of the actual problem at hand.

3 Errors and questionable issues

Overall, the paper is well-written and maintains mathematical rigor. Nevertheless, there are a few typographical errors and areas of concern that need addressing. These are categorized into minor and major technical issues as follows:

3.1 Minor technical issues

Kernel definition: The adopted kernel in the paper is defined as: $k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') \coloneqq \exp(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2l^2})$. This definition does not align with the standard form used in Gaussian process regression nor with the authors' code implementation, as the RBF kernel typically includes two hyperparameters: amplitude and lengthscale. The correct formulation should be: $k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') \coloneqq \sigma^2 \exp(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2l^2})$.

Posterior of causal Gaussian process: The posterior distribution of the causal Gaussian process presented in the paper is incorrect. This error may not affect the code implementation because it utilizes GPy, which handles the posterior calculations internally, thus obviating the need to construct it from scratch. Below is the original equation from the paper, with modifications highlighted in red. Note that the inversion of the covariance matrix, a critical operation that significantly influences the computational complexity of Gaussian processes, is omitted.

$$\begin{aligned} m_{s,t}(\mathbf{x} \mid \mathcal{D}^I_{s,t}) &= m_{s,t}(\mathbf{x}) + k_{s,t}(\mathbf{x}, \mathbf{X}^I)[k_{s,t}(\mathbf{X}^I, \mathbf{X}^I) + \sigma^2\mathbf{I}]^{-1}(\mathbf{Y}^I_{s,t} - m_{s,t}(\mathbf{X}^I) \text{ and} \\ k_{s,t}(\mathbf{x}, \mathbf{x}' \mid \mathcal{D}^I_{s,t}) &= k_{s,t}(\mathbf{x}, \mathbf{x}') - k_{s,t}(\mathbf{x}, \mathbf{X}^I)[k_{s,t}(\mathbf{X}^I, \mathbf{X}^I) + \sigma^2\mathbf{I}]^{-1}k_{s,t}(\mathbf{X}^I, \mathbf{x}'). \end{aligned} \tag{5}$$

3.2 Major technical issues

Modeling of $f_Y^{\mathbf{Y}}$ and $f_Y^{\mathbf{NY}}$ based on observational data: A fundamental prerequisite for the development of the DCBO methodology is the assumption of additivity in f_{Y_t}. This assumes that $Y_t = f_{Y_t}(pa(Y_t)) + \epsilon$, where $f_{Y_t}(pa(Y_t)) = f_Y^{\text{Y}}(Y_t^{\text{PT}}) + f_Y^{\text{NY}}(Y_t^{\text{PNT}})$. In the paper, both f_Y^{Y} and f_Y^{NY} are modeled as independent Gaussian processes using observational data. These two stochastic processes are crucial for establishing the prior mean and covariance functions in the DCBO model. However, relying solely on samples from the SEM would be inadequate for modeling these functions. Here we discuss the reasons for this limitation.

Consider the sequentially generated dataset $\{\mathbf{V}_t\}_{t=0}^{n_t}$, where each observation $\mathbf{V}_t = \{\mathbf{Z}_t, \mathbf{Y}_t\}$ captures measurements from every node in the causal graph $\mathcal{G}_t$ at time t. Here, $\mathbf{Y}_t$ denotes measurements of target variables, and $\mathbf{Z}_t$ represents measurements of all remaining manipulative and non-manipulative variables. Because every node can be modeled as a stochastic process of its parents, estimating the SEM from $\{\mathbf{V}_t\}_{t=0}^{n_t}$ is straightforward in principle.

However, focusing on the components f_Y^{Y} and f_Y^{NY} poses a deeper challenge. By construction, they appear in the additively decomposed form

$$f_{Y_t}(pa(Y_t)) = f_Y^{\text{Y}}(Y_t^{\text{PT}}) + f_Y^{\text{NY}}(Y_t^{\text{PNT}}), \tag{6}$$

where observations of Y_t^{PT} and Y_t^{PNT}, represented by $\mathbf{Y}_t^{\text{PT}}$ and $\mathbf{Y}_t^{\text{PNT}}$, emerge from past target variables $\{\mathbf{Y}_\tau\}_{\tau=0}^{t-1}$ and present-time manipulative/non-manipulative variables $\mathbf{Z}_t$, respectively. Although $\mathbf{Y}_t^{\text{PT}} \cup \mathbf{Y}_t^{\text{PNT}}$ collectively furnish all observations over $pa(Y_t)$, there are no direct measurements of either f_Y^{Y} or f_Y^{NY} themselves.

Proposition 1 (Non-Identifiability of Additive Components). *Let $\mathcal{X} \subseteq \mathbb{R}^m$, $\mathcal{Y} \subseteq \mathbb{R}^n$, and consider an unknown function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that admits an additive decomposition:*

$$f(x, t) = g(x) + h(y),$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $g : \mathcal{X} \to \mathbb{R}$ and $h : \mathcal{Y} \to \mathbb{R}$ are unknown sub-components. Suppose we only possess observational data $\{(x_i, y_i), f(x_i, y_i)\}_{i=1}^N$ but have no direct samples or constraints on $g(x)$ or $h(y)$ individually. Then, in general, $g(x)$ and $h(y)$ cannot be uniquely identified.

Proof: Consider any constant $c \in \mathbb{R}$. Let $g^\star(x) = g(x) + c$, and $h^\star(y) = h(y) - c$. It follows that $g^\star(x) + h^\star(y) = g(x) + h(y) = f(x, y)$. Hence, the observed mapping $f(x, y)$ remains unchanged despite shifting g and h by $\pm c$. Therefore, each constant c yields a valid decomposition $(g^\star, h^\star)$, there is infinitely many solutions that all produce the same $f(x, y)$. This establishes that solely relying on $\{(x_i, y_i), f(x_i, y_i)\}_{i=1}^N$ cannot isolate a unique pair (g, h).

In light of **Proposition 1**, which establishes the non-identifiability of two separate functions in an additive system $f(x, y) = g(x) + h(y)$ from observational data alone, the model components f_Y^{Y} and f_Y^{NY} cannot be individually recovered from the full observational dataset $\{\mathbf{V}_t\}_{t=0}^{n_t}$. Conventional observational data only provide the summed value $f_{Y_t}(pa(Y_t))$, and hence additional structure or external constraints (e.g., direct measurements, functional restrictions, or prior knowledge) must be imposed to achieve a unique decomposition into f_Y^{Y} and f_Y^{NY}.

The authors' code implementation treats the respective observations of f_Y^{Y} and f_Y^{NY} as if they were directly those of Y_{t-1} and Y_t, respectively, which is not exactly correct. As explained above, these two additive components cannot be individually identified by relying solely on samples drawn from the SEM; additional constraints are necessary. In particular, one viable approach might be to impose a one-step Markov structure on f_Y^{Y} by assuming:

$$f_Y^{\mathrm{Y}}(Y_t^{\mathrm{PT}}) = Y_{t-1} + \epsilon, \tag{7}$$

where ϵ is a stochastic perturbation. This assumption helps ensure that f_Y^{Y} can be more reliably disentangled based on observations from Y_{t-1}. In turn, identifying f_Y^{NY} requires analyzing the increment $Y_t - Y_{t-1}$, rather than Y_t alone.

Necessity of imposing assumption on additivity of f_{Y_t}: The additivity of f_{Y_t} is a core assumption in DCBO method, underpinning the conclusion in Eq. 2. This assumption allows us to separate the temporal transition effects on the target variable. For each minimal intervention set, this separation translates into a shift, as the inputs of f_Y^{Y} do not intersect with $\mathbf{x}_{s,t}$. Based on this conclusion, all causal Gaussian process priors in this paper are established. A critical question arises: while this method provides a robust way to find informative priors, what advantages does it offer over the direct sampling approach over $Y_t|(\mathrm{do}(\mathbf{X}_{s,t} = \mathbf{x}_{s,t}), \mathbb{1}_{t>0} \cdot I_{0:t-1})$ introduced in Causal Bayesian Optimization?

Fundamentally, DCBO performs causal Bayesian optimization at each time step during sequential optimization. The priors for the causal Gaussian process can be obtained either through direct sampling as described in causal Bayesian optimization or through the secondary conclusion derived from Eq. 2.

Direct sampling involves first estimating the SEM based on observational data. Then, under specified intervention conditions, that is, $(\mathrm{do}(\mathbf{X}_{s,t} = \mathbf{x}_{s,t}), \mathbb{1}_{t>0} \cdot I_{0:t-1})$, sampling from the estimated SEM yields the mean and standard deviation of Y_t, which serve as priors for the causal Gaussian process.

In contrast, leveraging the conclusion from Eq. 2 involves estimating the SEM, f_Y^{Y} and f_Y^{NY} from observational data. At a given time step t, the optimal target variable within $pa(Y_t)$ is input into f_Y^{Y} to compute the output distribution. Simultaneously, the distribution of f_Y^{NY} under inputs $\mathbf{x}^{\mathrm{PY}}, \mathbf{i}^{\mathrm{PY}}, \mathbf{w}$ is computed, requiring sampling from the estimated SEM to obtain $\mathbf{w}|(\mathrm{do}(\mathbf{X}_{s,t} = \mathbf{x}), I_{0:t-1})$. Eventually, the output distributions from f_Y^{Y} and f_Y^{NY} are combined to derive the priors for the causal Gaussian process.

Comparing these two strategies for obtaining priors, both aim to estimate $Y_t|(\mathrm{do}(\mathbf{X}_{s,t} = \mathbf{x}_{s,t}), \mathbb{1}_{t>0} \cdot I_{0:t-1})$ based on observational data. Under equivalent conditions, the direct sampling approach from causal Bayesian optimization appears advantageous due to its lower computational cost, simpler steps, and the absence of the need to estimate f_Y^{Y} and f_Y^{NY}. Therefore, the assumption of additivity of f_{Y_t} may seem superfluous.

Parameters that define the synthetic experiments: The first three synthetic experiments are based on the same SEM setting, given by:

$$\begin{aligned} X_t &= X_{t-1}\mathbb{1}_{t>0} + \epsilon_X, \\ Z_t &= Z_{t-1}\mathbb{1}_{t>0} + \exp(-X_t) + \epsilon_Z, \\ Y_t &= Y_{t-1}\mathbb{1}_{t>0} - \exp(-Z_t/20) + \cos(Z_t) + \epsilon_Y, \end{aligned} \tag{8}$$

where $\epsilon_i \sim \mathcal{N}(m_i, \sigma_i)$ for $i \in \{X, Y, Z\}$ are considered as Gaussian random noise, and $\mathbb{1}_{t>0}$ is an indicator function that equal to one when $t > 0$ and zero otherwise. The parameters that define these synthetic experiments include the intervention domain for each manipulative variable, and control parameters m_i and σ_i for Gaussian random noise.

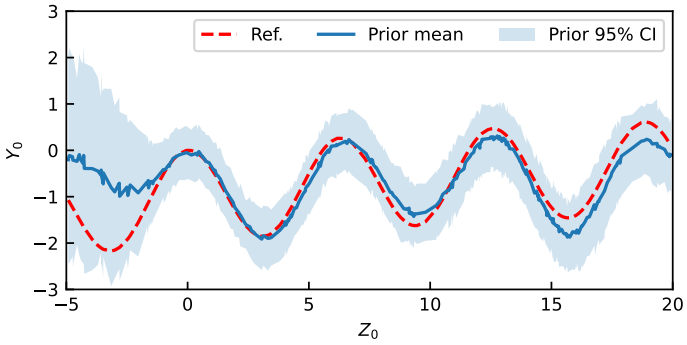

Figure 1: Prior for causal Gaussian process $f_{Z,0} = Y_0|\text{do}(Z_0)$.

By definition, the form $Z_t = Z_{t-1}\mathbb{1}_{t>0} + \exp(-X_t) + \epsilon_Z$ suggests that Z_t should remain strictly positive in the noise-free limit (since $\exp(-X_t)$ is always positive). Even with the noise contamination ϵ_Z, unless it is large and negative, we generally expect most Z_t values to remain above zero when the stochastic process is allowed to run sequentially for multiple time steps. Indeed, if the variance of ϵ_Z is not excessively large, observing negative values of Z_t in practice becomes rare. However, the intervention domain in the paper is $D(Z_t) = [-5.0, 20.0]$, thus permitting manipulations in the interval $[-5.0, 0)$. This raises two issues. First, it contradicts the original assumption that Z_t is inherently nonnegative. Second, because the defined SEM rarely generates negative data for Z_t, the function mapping Z_t to Y_t (i.e., f_Y^{NY}) will have limited or even no observational data in that negative subdomain, resulting in a potentially weak or uninformative prior for the causal Gaussian process. From a practical standpoint, such an intervention domain can lead to confusion during both analysis and optimization. If it truly is necessary that Z_t take negative values, the SEM itself should be modified or re-justified to accommodate them. Otherwise, as shown in Figure 1, the scarcity of negative observations for Z_t can undermine the reliability of the prior for the causal Gaussian process in $[-5.0, 0)$, even though this prior is derived from a setting with low observational noise. Notably, in the very first step of dynamic causal Bayesian optimization, the algorithm might identify an apparent optimum in this inadequately sampled region, ultimately compromising the interpretability of the results.

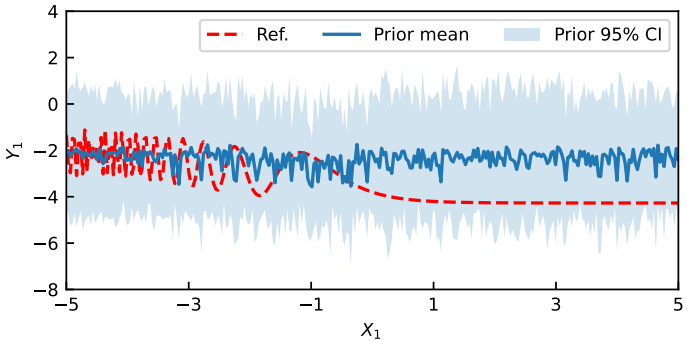

Figure 2: Prior for causal Gaussian process $f_{X,1} = Y_1|(\text{do}(X_1), I_0)$.

Another potential issue is that the noise parameters set in the examples may be unreasonable. Excessive standard deviations render the prior for $Y_t|(\text{do}(\mathbf{X}_{s,t} = \mathbf{x}_{s,t}), \mathbb{1}_{t>0} \cdot I_{0:t-1})$ estimated from observational data non-informative. When the prior is non-informative, the causal Gaussian process cannot effectively leverage its advantages. As illustrated in Figure 2, the prior for $Y_1|(\text{do}(X_1), I_0)$ derived from the noise parameters in synthetic experiment 1 is non-informative. Under these circumstances, even though the global optimum can be found after several iterations, this outcome is

attributable to the inherent properties of Bayesian optimization rather than the advantages of the method proposed in this paper.

4 Areas lacking clarity and derivations

Theorem 1 in the paper proposed that the intervention function $f_{s,t}(\mathbf{x}) : D(\mathbf{X}_{s,t}) \to \mathbb{R}$ can be decomposed into two parts, as shown in Eq. 2, where the second part is challenging to evaluate because its expectation is taken with respect to a distribution conditioned on do-operations. To address this difficulty, the authors introduce **Corollary 1**, which states:

$$\mathbb{E}_{p(\mathbf{w}|\text{do}(\mathbf{X}_{s,t}=\mathbf{x}),I_{0:t-1})}\left[f_Y^{\text{NY}}(\mathbf{x}^{\text{PY}},\mathbf{i}^{\text{PY}},\mathbf{w})\right] = \mathbb{E}_{p(\mathbf{U}_{0:t})}\left[f_Y^{\text{NY}}(\mathbf{x}^{\text{PY}},\mathbf{i}^{\text{PY}},\{C(W)\}_{W\in\mathbf{W}})\right]. \tag{9}$$

where $p(\mathbf{U}_{0:t})$ is the distribution of the exogenous variables up to time t and $C(W)$ is defined via:

$$C(W) = \begin{cases} f_W(\mathbf{u}_W, \mathbf{x}^{\text{PW}}, \mathbf{i}^{\text{PW}}) & \text{if} \quad R = \emptyset \\ f_W(\mathbf{u}_W, \mathbf{x}^{\text{PW}}, \mathbf{i}^{\text{PW}}, r) & \text{if} \quad R \subseteq \mathbf{X}_{s,t} \cup I^V_{0:t-1} \\ f_W(\mathbf{u}_W, \mathbf{x}^{\text{PW}}, \mathbf{i}^{\text{PW}}, C(R)) & \text{if} \quad R \not\subseteq \mathbf{X}_{s,t} \cup I^V_{0:t-1} \end{cases} \tag{10}$$

where f_W is the structural function for W in the SEM, $\mathbf{u}_W$ is the set of exogenous variables feeding into W. $\mathbf{x}^{\text{PW}}$ and $\mathbf{i}^{\text{PW}}$ are the values corresponding to $\mathbf{x}^{\text{PW}}_{s,t}$ and $I^{\text{PW}}_{0:t-1}$ which in turn represents the subset of variables in $\mathbf{X}_{s,t}$ and $I^V_{0:t-1}$ that are parents of W. Finally r is the value of $R = pa(W)\backslash(\mathbf{X}^{\text{PW}}_{s,t} \cup I^{\text{PW}}_{0:t-1})$. Note that the red-colored "W" is a corrected error from the original context. Below, we provide a more explicit derivation for **Corollary 1**.

In general, each non-intervened variable $W \in \mathbf{W}$ is governed by a structural equation in the SEM, given by:

$$W = f_W(pa(W)) + \epsilon_W, \tag{11}$$

with ϵ_W is an exogenous noise term. If some variables in $pa(W)$ are also non-intervened, they could be recursively expanded as well based on Eq. 11. Eq. (10) compactly encodes this procedure by "unrolling" any non-intervened parent into its own structural equation. Essentially, there are three conditions covered by Eq. 10: (i) If W has no other non-intervened parents ($R = \emptyset$), expand W directly via its structural function. (ii) If all parents in R are themselves set by interventions, substitute these values directly. (iii) Otherwise, keep recursing by calling $C(R)$. By replacing each $\mathbf{w}$ by $\{C(W)\}_{W\in\mathbf{W}}$, we eliminate the do operator in the distribution:

$$p(\mathbf{w}|\text{do}(\mathbf{X}_{s,t} = \mathbf{x}), I_{0:t-1}) = p(\{C(W)\}_{W\in\mathbf{W}}|\text{do}(\mathbf{X}_{s,t} = \mathbf{x}), I_{0:t-1}) = p(\mathbf{U}_{0:t-1}). \tag{12}$$

Here, the last equality follows that every non-manipulated variable is determined by exogenous noise plus possibly manipulated parents. Hence, the randomness is carried solely by $\mathbf{U}_{0:t-1}$. Substituting back into the expectation $\mathbb{E}_{p(\mathbf{w}|\text{do}(\mathbf{X}_{s,t}=\mathbf{x}),I_{0:t-1})}\left[f_Y^{\text{NY}}(\mathbf{x}^{\text{PY}},\mathbf{i}^{\text{PY}},\mathbf{w})\right]$, we obtain $\mathbb{E}_{p(\mathbf{U}_{0:t})}\left[f_Y^{\text{NY}}(\mathbf{x}^{\text{PY}},\mathbf{i}^{\text{PY}},\{C(W)\}_{W\in\mathbf{W}})\right]$, which completes the derivation of **Corollary 1**. In essence, every non-intervened parent is replaced by a structural function of exogenous noise and known intervention values, making the do operator unnecessary in the conditional probability.

5 Code evaluation and replication

5.1 Critical code components omitted in the paper

Modeling of all SEM-related functions using Gaussian processes: Constructing Gaussian process regression models for these functions based on observational data provides a foundation in subsequent Bayesian optimization steps. The challenge lies in efficiently implementing these function models in the program. In the authors' code, a method named `fit_arcs` is created, which takes three inputs: a `MultiDiGraph` object, a dictionary representing the required data, and a boolean value to determine which edge-type is being processed (emissions or transitions). Emissions represent causality within each time slice, while transitions denote causality propagated over time.

The output of `fit_arcs` is the SEM-related functions for all four Bayesian optimization methods as mentioned in the paper. Within the `fit_arcs` method, the adjacency matrix A for the current graph $\mathcal{G}_t$ can be obtained from the callable `MultiDiGraph`. By distinguishing between emissions and transitions, A is decomposed into A_e (emission adjacency

matrix) and A_t (transition adjacency matrix). Since the subsequent steps for modeling emission and transition functions are similar, we will illustrate the process using emissions as an example. First, nodes corresponding to rows in A_e with more than one non-zero element are identified as fork nodes. The mappings from these fork nodes to their child nodes is initially modeled using Gaussian processes. Next, nodes corresponding to rows in A_e with all zero elements are considered. These nodes' generation mechanisms do not depend on other nodes and are modeled using random variables. The remaining nodes, which have exactly one non-zero element in their corresponding rows in A_e, are then modeled. Finally, nodes with multiple inputs, as indicated by A_e, are modeled to capture the multivariate mapping.

The method `fit_arcs` enables the establishment of all SEM-related functions required for the four Bayesian optimization approaches as adopted in the paper. The design of this method exhibits two notable features. First, a binary matrix `edge_fit_track_mat` is initialized to keep track of which edges have been fitted. This ensures that all edges are accounted for and fitted exactly once. Second, nodes with multiple children (fork nodes) are identified and processed first, ensuring that complex dependencies are handled early in the process.

Sequential sampling from the true and estimated SEM: The code stores all functions for true and estimated functions in an `OrderedDict`. During sequential sampling, the algorithm can generate samples in batch by simply traversing each element in the `OrderedDict` in order. When considering specific interventions, the algorithm can skip the `OrderedDict` to generate data at the point of the intervention and directly use the specified intervention values. Subsequent computations follow this principle, enabling the generation of intervened samples.

Furthermore, to simplify the experimental process, the authos' code adjusts the true SEM to be noise-free when generating intervention data for $\mathbf{X}_{s,t}$. This adjustment allows the representation of $\mathbb{E}\left[Y_t|\text{do}(\mathbf{X}_{s,t} = \mathbf{x}_{s,t}), \mathbb{1}_{t>0} \cdot I_{0:t-1}\right]$, with just a single sample, eliminating the need for extensive Monte Carlo sampling and subsequent mathematical expectation calculations. This simplification is valid, as demonstrated in synthetic experiment 1, only if every noise term in the SEM has zero mean. Specifically, if $\mathbb{E}(\epsilon_V) = 0$ for each noise term ϵ_V, then removing ϵ_V does not alter the expected value of the corresponding variable V. In that case, the true function $V = f_V(pa(V)) + \epsilon_V$ implies $\mathbb{E}(V) = f_V(pa(V)) + \mathbb{E}(\epsilon_V) = f_V(pa(V))$. Consequently, all nodes including Y_t in the causal graph accurately preserves the original expectations under the specified interventions. However, if any noise term has a nonzero mean, omitting it introduces a systematic bias in the intervened data, which can lead to invalid inferences about intervention effects.

5.2 Insights from code analysis and replication experience

Computational efficiency: The evaluation of prior mean and covariance functions is computationally demanding because it requires a large number of samples based on the estimated SEM, which is represented by a series of Gaussian processes. This process consumes most of the computational resources and time required for computing the posterior of the causal Gaussian processes, as given in Eq. 2. Upon analyzing the authors' code and our replication results, two key strategies can be employed to optimize the algorithm's efficiency. First, parallelize the computation of prior mean and covariance functions (as mentioned by the authors in the comments for `mean_function_internal` and `variance_function_internal`, where they suggest parallelizing all sampling functions). Second, avoid redundant computations during the computation of the posterior. While existing packages such as GPy and TensorFlow-Probability already provide internal algorithmic optimizations, it is crucial to pay attention to this aspect even after defining custom causal kernels.

Accuracy of Gaussian process models: In the framework of the DCBO methodology, numerous Gaussian process models are established. These include many SEM-related functions as well as causal Gaussian processes. Generally, the optimization of Gaussian process hyperparameters involves a non-convex objective function, specifically the negative log marginal likelihood, when considering observational noise variance. If the initial values of the hyperparameters are poorly chosen, the optimization may converge to local minima, leading to inaccurate predictions by the Gaussian process models. In practical applications, any issues with a single Gaussian process model can significantly impact the prediction of optimal solutions by the DCBO algorithm. These negative impacts may include convergence to local optima or a substantial increase in the cost of reaching the global optimum. Therefore, it is essential to introduce a mechanism to ensure the correctness of the Gaussian process models. Although manual analysis of each constructed

Gaussian process model can be a conservative strategy, it becomes increasingly difficult as the complexity of the causal relationships increases. In the authors' code, they employ the `optimize_restarts` function provided by GPy, which performs random restarts of the model to find the best seen solution. Additionally, the TensorFlow Probability framework offers the ability to marginalize hyperparameters using Hamiltonian Monte Carlo, which can mitigate sensitivity to initial hyperparameter values.

6 Potential applications in banking: Portfolio optimization in a dynamic market change

Portfolio optimization is a fundamental task in banking and finance, focusing on the strategic selection of an optimal mix of assets to maximize returns for a given level of risk, or conversely, to minimize risk for a desired level of expected return [7]. Indeed, Bayesian optimization has been a practical tool for numerous portfolio optimization problems [8]. However, DCBO is expected to transformatively outperform traditional Bayesian optimization because it introduces two more important things that could better describe the market, i.e., the market dynamics and causality of different factors in a market, which eventually will lead to a more informed proactive portfolio management. The detailed implementation of DCBO is described below.

First, potential causal relationships among various market indicators across both temporal and asset dimensions are identified. These indicators may include interest rates, inflation rates, GDP growth, geopolitical events, and more. A causal graph is then constructed in which nodes represent both manipulable and non-manipulable market factors. This graph serves as both a visual and analytical tool to understand and predict the effects of changes in one variable on others. Next, historical data on the identified factors are collected, and Gaussian processes are built based on these observations to construct an initialized SEM. If real-time data are available, the SEM is updated sequentially. Using this framework, the prior for the Causal Gaussian Processes, which describe the intervention effects of various explorations on market returns, is constructed by sampling from the SEM. Decisions can then be made regarding where to intervene next to achieve the highest expected return improvement. Ultimately, the optimal portfolio allocation strategy is determined.

References

[1] Virginia Aglietti, Neil Dhir, Javier González, and Theodoros Damoulas. Dynamic causal Bayesian optimization. In *Advances in Neural Information Processing Systems*, pages 10549–10560, 2021.

[2] Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal Bayesian optimization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 3155–3164, 2020.

[3] Jonas Peters, Stefan Bauer, and Niklas Pfister. Causal models for dynamical systems, 2020.

[4] Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems*, pages 2573–2583, 2018.

[5] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[6] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[7] Daniel P. Palomar. *Portfolio Optimization: Theory and Application*. Cambridge University Press, 2024.

[8] Matthew Hoffman, Eric Brochu, Nando De Freitas, et al. Portfolio allocation for Bayesian optimization. In *UAI*, pages 327–336, 2011.