High-Dimensional Causal Bayesian Optimization

Yupeng Wu[a,1], Weiye Wang[a,1], Yangwenhui Zhang[a], Mingjia Li[a], Yuanhao Liu[a], Hong Qian[a,*] and Aimin Zhou[a]

[a]Shanghai Institute of AI for Education and School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

Abstract. Causal global optimization (CGO) aims to complete optimization tasks through causal inference. In the high-dimensional CGO problems, traditional causal Bayesian optimization (CBO) methods struggle with the curse of dimensionality attributed to the number of variables in the causal graph, and scale inconsistency among Gaussian Process (GP) models. These issues limit the application of CBO in domains requiring optimization over large causal graphs. To address these limitations, this paper proposes a high-dimensional causal Bayesian optimization (HCBO) algorithm. To address the curse of dimensionality, HCBO introduces a submodularity indicator for variable subsets through the concept of causal intrinsic dimensionality (CID). It then uses the submodular optimization algorithm to find approximations of CID within polynomial sample complexity. Theoretically, we disclose a sufficient condition for CID's existence. To address the issue of scale inconsistency among GP models, HCBO introduces a scale-normalized scoring function, ensuring stable identification of the optimal GP model corresponding to CID for intervention. Extensive experiments are conducted on high-dimensional synthetic and real-world tasks, i.e., coral ecology and health. The existence of CID is verified across the datasets of all tasks. HCBO achieves state-of-the-art performance in CGO problems and can handle causal graphs at a scale 10 times larger than that manageable by previous CBO methods.

1 Introduction

Black-box optimization [8, 24] has significant applications in statistics and industry. It treats the objective function as a black-box entity, conducting optimization without relying on gradients or problem-specific information. This approach models the objective function in a unified form $f : \mathcal{X} \subset \mathbb{R}^D \to \mathbb{R}$. Widely-adopted black-box optimization techniques can be broadly categorized into three types [31], including Lipschitzian-based partitioning methods, population-based methods, and model-based methods. Bayesian optimization (BO), the representative model-based black-box optimization method, is successfully applied in machine learning [34], reinforcement learning [29], and scientific computing [33]. The classical BO methods employ Gaussian Process (GP) as the surrogate model to approximate the objective function f. The global optimization is achieved by maximizing an acquisition function derived from GP, which determines the next solutions for evaluation. However, existing work [1, 2, 7, 3, 37, 16] have observed that BO does not achieve optimal performance in optimization scenarios where the causal graph [26] of the problem is known. This observation has led to the proposal of causal global optimization (CGO) problem and causal Bayesian optimization (CBO) method [1].

CGO offers two significant advantages: it reduces dimensions based on the causal graph and discovers that optimizing some variables can achieve global optimality, whereas optimizing the entire set cannot [1]. Here are two easily understandable real-world examples that illustrate the practicality and importance of CGO. In agriculture [23], dozens of heavy metals in the soil are statistically correlated with the optimization target, e.g., plant growth. However, through causal inference, the complex effects of heavy metals can be reduced into effects on a limited number of variables such as the pH value of the soil, significantly simplifying the optimization problem. In biology [32], intermediate products cannot be directly intervened to the optimal value due to limited control domains and must rely on other key chemical raw materials to react. Even, due to the order of reactions, controlling all intermediate products (all variables) results in local optima while controlling a subset of variables can yield better outcome. In summary, the incorporation of causal priors aids in the resolution of the global optimization problems.

The requirement for a known causal graph in CGO problems is not an insurmountable barrier. In fact, in industrial, medical and biological fields [9, 39, 14, 32], expert-annotated causal graphs, which qualitatively describe causation between optimization variables, are often accessible. Additionally, CEO [7] provides an optimization approach for CGO problems where the causal graph is unknown. The utilization of these causal graphs enables CBO methods to significantly enhance the optimization process. This includes benefits of lower optimization cost and superior target value. Specifically, the causal graph enables CBO methods to concentrate only on a subset of optimization variables, simplifying the problem while lowering the cost of intervention. Furthermore, by clarifying the complex interrelations between variables, causal graphs facilitate a more direct approach to the global optimum. While CBO has shown promising results, it is noted that no work has yet solved high-dimensional CGO problems.

The classical CBO method [1] can handle small-scale CGO problems effectively, but its efficacy markedly declines in high-dimensional CGO problems. This downturn in performance is primarily attributed to the curse of dimensionality caused by the number of variables in the causal graph and the scale inconsistency among multiple GP models. Specifically, the traditional CBO method determines an exploration set (ES) for intervention, comprising multiple variable subsets, and constructs a corresponding number of GP models. After balancing exploration and exploitation, CBO method selects a promising GP model by the acquisition function for intervention. The curse of dimensionality refers to the exponential growth of the ES scale

* Corresponding Author. Email: hqian@cs.ecnu.edu.cn.

[1] Equal contribution.

with the number of variables in the causal graph. Moreover, the sample complexity of the algorithm for searching ES is also exponential. Scale inconsistency in GP models presents challenges in balancing exploration and exploitation. The term "scale" refers to the domain of target variable modeled by each GP model. To overcome these issues, reducing the scale of ES and eliminating the scale inconsistency are the primary research focus of this work.

To this end, this paper presents the first algorithm to address the high-dimensional CGO problems, termed high-dimensional causal Bayesian optimization (HCBO) algorithm. The contribution of this work is three folds. (I) A state-of-the-art (SOTA) method for high-dimensional CGO problem. Compared to CBO methods, HCBO can handle causal graphs of sizes up to 200, a significant increase from about 10. Compared to high-dimensional BO methods, HCBO achieves SOTA results in terms of cost and target value indicators by leveraging causal graph. (II) An ES efficient search strategy named ES-ESS, based on a proposed submodular indicator, for HCBO to effectively overcome the curse of dimensionality attributed to the size of causal graph. When the causal intrinsic dimensionality (CID) exists, compared to CBO methods, HCBO can identify a polynomial scale, high-quality approximation for an exponential scale ES. Unlike high-dimensional BO methods that sometimes rely on a random approach to find intrinsic dimensions [41, 30], HCBO automatically determines the CID based on the submodular indicator. Additionally, Theorem 2 provides a sufficient but not necessary condition for CID's existence. (III) A scale-normalized intervention set score function named ISSF for HCBO to select optimal GP model, unaffected by scale inconsistency. As the problem scale increases, HCBO can stably find the globally optimal GP model among the polynomial-scale models.

In the subsequent sections, we respectively review the related work, present the preliminaries and problem statement, introduce the proposed HCBO, show the experimental results and analysis, and finally conclude the paper. The code and Appendix are available at https://github.com/ANormalMan12/HCBO.

2 Related Work

Causal Bayesian Optimization. Building upon the foundations of BO, CBO methodologies [1, 2, 7, 3, 37, 16] leverage causal inference to perform optimization tasks. They have demonstrated strong empirical performance across various domains such as manufacturing, ecology, communication, medicine and healthcare, system biology, operational research, and social science. The seminal work [1] formalized the modeling of CBO. Subsequent developments like DCBO [2], CEO [7], and cCBO [3] made targeted progress in dynamical, DAG-unknown, and constrained causal systems, respectively. MCBO [37] introduced the first method with sublinear cumulative regret convergence guarantees for CBO, enhancing CBO's empirical performance. Additionally, fCBO [16] and MCBO [37] explored the impact of contextual interventions on CGO problems, showcasing the significant benefits of leveraging causal prior for optimization. However, these efforts have primarily focused on low-dimensional CGO scenarios with causal graphs of no more than 10 optimization variables. To our knowledge, this paper is the first to systematically address high-dimensional CGO problems, supporting the causal graphs of sizes up to 200 and achieving SOTA performance.

High-Dimensional Bayesian Optimization. High-dimensional Bayesian optimization (HBO) methods, which do not utilize causal priors, are ill-suited for solving high-dimensional CGO problems. HBO methods often highlight the concept of intrinsic dimensions as a method for managing the curse of dimensionality. This segment covers linear embedding methods, non-linear embedding methods, and variable selection methods. REMBO [41] introduces a technique wherein a randomly generated embedding matrix connects a low-dimensional subspace with the original optimization space, thereby sidestepping the curse of dimensionality through optimization within this subspace. Subsequent efforts in random embedding [5, 30, 6, 25, 21] focus on enhancing the embedding quality to ensure that the low-dimensional subspace encompasses the global optima. Meanwhile, other approaches, such as [4] and [44], endeavor to learn this embedding matrix. Unlike linear embedding methods, non-linear embedding methods leverage machine learning to autonomously devise a map that supplants the embedding matrix, with variational autoencoders emerging as the most frequently employed technology [15, 40]. In the realm of variable selection methods [22, 35], it is posited that only a subset of the input dimensions impact the objective function value, with the rest having no effect, a notion akin to the CGO problem setting. Dropout BO [22] randomly selects d dimensions to construct the effective dimension sets, thereby obtaining a cumulative regret convergence rate that is dependent on d. Conversely, MCTS-VS [35] utilizes the Monte-Carlo tree search method to identify the optimal effective dimension set, albeit requiring significant effort to assess different dimension sets. In the CGO problem, the known causal graph is underutilized by traditional HBO methods, making the identification of intrinsic dimensions inefficient and challenging.

3 Preliminaries and Problem Statement

3.1 Structural Causal Models

A structural causal model $\mathcal{M}$ [27, 28] is a type of causal modeling framework which can be described by a known directed acyclic graph $\mathcal{G}$ and a four-tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$, where $\boldsymbol{U}$ denotes a collection of exogenous variables, the values of which are determined by factors external to $\mathcal{M}$. $\boldsymbol{V}$ comprises a set of endogenous observable variables whose values are influenced by the variables in $\mathcal{M}$. Variables in $\boldsymbol{V}$ can be classified into three categories: $\boldsymbol{C}$, $\boldsymbol{I}$ and Y. Background variables $\boldsymbol{C}$ are immutable and cannot be manipulated. Treatment variables $\boldsymbol{I}$ are manipulative and can be assigned specific values. The output variable Y denotes the agent's outcome of interest, whose variations are related to changes in $\boldsymbol{C}$ and $\boldsymbol{I}$. $\boldsymbol{F}$ is a collection of structural equations $(f_1, \ldots, f_V)$, delineating the functional relationships among variables. Each f_i maps from the corresponding domains of $U \cup Pa(V)$ to V, where $U \subseteq \boldsymbol{U}$ and $Pa(V)$ is the set of parent variables of the single variable V in $\mathcal{G}$. $P(\boldsymbol{U}) = \prod_{U \in \boldsymbol{U}} p(U)$ is the probability distribution of the set of exogenous variables $\boldsymbol{U}$.

3.2 Interventions & Do-Calculus

Intervention and do-calculus [27] are both important concepts in the theory of the structural causal model. Intervention is one of the most important operations which can change the probability distribution of the variables in causal graph $\mathcal{G}$, while do-calculus allows us to calculate the distribution after intervention from the observational data. The intervention and do-calculus are detailed below.

An intervention that assigns specific values $\boldsymbol{x}$ to the treatment variables $\boldsymbol{X} \subseteq \boldsymbol{I}$ is represented as $do(\boldsymbol{X} = \boldsymbol{x})$, corresponding to modifying $\mathcal{M}$ by substituting the functional relations $f(Pa(\boldsymbol{X}), U_{\boldsymbol{X}})$ with $\boldsymbol{x}$. $\boldsymbol{X}$ is referred to the intervention set. The interventional distribution for two disjoint sets $\boldsymbol{X}$ and Y is denoted as $P(Y|do(\boldsymbol{X} = \boldsymbol{x}))$, which signifies the distribution of Y resulting from intervening on $\boldsymbol{X}$ while fixing its values to $\boldsymbol{x}$. $P(Y|do(\boldsymbol{X} = \boldsymbol{x}))$ can be approximated

by getting a Monte Carlo estimate. The interventional expectation of do-calculus is denoted as $\mathbb{E}\left[Y|do(\boldsymbol{X}=\boldsymbol{x})\right]$.

3.3 Causal Intrinsic Dimensionality (CID)

In the context of HBO problems, standard BO methods quickly degrade due to the curse of dimensionality [10, 11]. The assumption of intrinsic dimensionality is a way for HBO methods to mitigate this curse, suggesting that high-dimensional optimization problems have a low-dimensional subspace where optimizing within this subspace can lead to the optimal solution [41]. Similarly, the assumption of CID holds potential for solving high-dimensional CGO problems. Def. 1 provides a formal definition of causal intrinsic dimensions $\boldsymbol{X}_e^*$, requiring the CID $d_e = |\boldsymbol{X}_e|$ strictly less than the full dimensions $|\boldsymbol{I}|$. Fig.1 (a) demonstrates an example of CID $d_e = 1$ in $\mathcal{G}$ with $\boldsymbol{I} = 100$.

Definition 1 (Causal Intrinsic Dimensionality, CID)**.** *For CGO problem, it is said to have an **effective intervention set** $\boldsymbol{X}_e \subset \boldsymbol{I}$. $\boldsymbol{X}_e$ is the subset of $\boldsymbol{I}$ but can achieve the global optima. Let $\mathbb{X}_e$ denote the collection of all effective intervention sets, and $Co(\boldsymbol{X})$ denote the intervention cost of the intervention set $\boldsymbol{X}$. We define the **causal intrinsic intervention set** $\boldsymbol{X}_e^*$ as the effective intervention set with the lowest intervention cost, and the **causal intrinsic dimensionality** d_e as the dimension of $\boldsymbol{X}_e^*$, where $\boldsymbol{X}_e^* = \arg\min_{\boldsymbol{X}_e \in \mathbb{X}_e} Co(\boldsymbol{X}_e)$.*

3.4 High-Dimensional Causal Bayesian Optimization

Addressing high-dimensional optimization problems within a causal framework presents a novel and effective approach, leveraging the causal relationships between variables. In high-dimensional CGO problems, given $\mathcal{G}$ and $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U})\rangle$, if the CID $d_e < |\boldsymbol{I}|$ exists, the goal is to find the causal intrinsic intervention set $\boldsymbol{X}_e^\star$ and the optimal intervention values $\boldsymbol{x}^\star$ to optimize the target outcome Y. Formally, it can be written as follows:

$$\boldsymbol{X}_e^\star, \boldsymbol{x}^\star = \underset{\boldsymbol{X}\in\mathcal{P}(\boldsymbol{I}), \boldsymbol{x}\in D(\boldsymbol{X})}{\arg\max} \mathbb{E}\left[Y|do(\boldsymbol{X}=\boldsymbol{x})\right], \tag{1}$$

where $\mathcal{P}(\cdot)$ is the power set and $D(\boldsymbol{X}) = \times_{I\in\boldsymbol{X}} D(I)$ with $D(I)$ denoting the intervention domain of a single treatment variable I. Note that in practical applications, there are cases where the optimal variable values lie outside the intervention domain, necessitating intervention on the ancestor variables to locate the global optima [1]. This is why CBO methods can identify superior optimal values, compared to general optimization methods.

To overcome the curse of dimensionality, CBO methods compress the search space of $\boldsymbol{X}_e^*$ from $\mathcal{P}(\boldsymbol{I})$, into a smaller exploration set (ES). Minimal intervention set MIS($\mathcal{M}$) is a common technique to search ES [20]. MIS($\mathcal{M}$) is a precise method that ensures ES encompasses $\boldsymbol{X}_e^*$. Standard CBO method [1] can solve Eq. (1) with MIS($\mathcal{M}$) under small scale $\mathcal{G}$, but it still suffers performance degradation in high-dimensional CGO problems. As $\mathcal{G}$ scaling up, the scale of ES obtained from MIS($\mathcal{M}$) grows exponentially. Concurrently, an exponential scale ES exacerbates the issue of scale inconsistency mentioned before. This paper presents a polynomial-scale approximation of ES and an intervention set score function to eliminate scale inconsistency.

4 Methodology

This section introduces the high-dimensional causal Bayesian optimization (HCBO) algorithm capable of solving Eq. (1) in high-dimensional scenarios. HCBO follows the CBO framework [1], with

Algorithm 1 Calculation of Causal Coverage of $\boldsymbol{X}$

Input: Causal graph $\mathcal{G}$, Endogenous observable variables $\boldsymbol{V}$ and corresponding weights $\boldsymbol{w}$, Target variable Y, Intervention set $\boldsymbol{X}$
Output: $\boldsymbol{X}$'s causal coverage $c_{\boldsymbol{X}}$

1: $c_{V|\boldsymbol{X}} \leftarrow 0, \forall V \in \boldsymbol{V}$% $c_{V|\boldsymbol{X}}$ is the conditional causal coverage
2: **for** $V \in \boldsymbol{V} \cup \{Y\}$ in a topological order of $\mathcal{G}$ **do**
3: **if** $V \in \boldsymbol{X}$ **then**
4: $c_{V|\boldsymbol{X}} \leftarrow w_V$
5: **else**
6: $c_{V|\boldsymbol{X}} \leftarrow \sum_{P\in Pa(V)} c_{P|\boldsymbol{X}}/|Pa(V)|$
7: **end if**
8: **end for**
9: **return** $c_{\boldsymbol{X}} \leftarrow c_{Y|\boldsymbol{X}}$

significant contributions made in two parts: Part (i) is an ES effective search strategy named ES-ESS with polynomial sample complexity for overcoming the curse of dimensionality attributed to the size of $\mathcal{G}$. ES-ESS can limit the scale of ES and the number of corresponding GP models, accelerating the convergence of HCBO. Part (ii) is a scale-normalized intervention set score function (ISSF) for overcoming the issue of scale inconsistency. ISSF can identify the causal intrinsic dimensionality set $\boldsymbol{X}_e^*$, speeding up the convergence of HCBO. Fig.1 shows the HCBO framework.

4.1 ES Efficient Search Strategy (ES-ESS)

ES-ESS aims to identify a smaller ES from $\mathcal{P}(\boldsymbol{I})$. CBO methods use the precise MIS($\mathcal{M}$) to obtain ES. MIS($\mathcal{M}$) guarantees the $\boldsymbol{X}_e^*$, but MIS($\mathcal{M}$) scales exponentially with the size of $\mathcal{G}$. ES-ESS can obtain a polynomial approximation of MIS($\mathcal{M}$), termed as the efficient causal coverage intervention set ECCIS($\mathcal{M}$). Specifically, ES-ESS employs a *causal coverage* indicator with submodularity, to score intervention sets within $\mathcal{P}(\boldsymbol{I})$. ES-ESS then utilizes a submodular optimization algorithm to search for high-scoring intervention sets to form ECCIS($\mathcal{M}$). ECCIS($\mathcal{M}$) represents a high-quality approximation of MIS($\mathcal{M}$), better suited for high-dimensional CGO problems.

4.1.1 Causal Coverage Indicator

The intervention set $\boldsymbol{X} \subseteq \boldsymbol{I}$ is the subset of the ancestor variables of Y. Different $\boldsymbol{X}$ lead to different optimal interventions on Y, denoted as $\mathbb{E}[Y|do(\boldsymbol{X}=\boldsymbol{x}^*)]$ [1], where $\boldsymbol{x}^*$ refers to the optimal intervention values within the domain $D(\boldsymbol{X})$. In causal theory, unlike causal effects [23] that describe the average magnitude of changes in $\boldsymbol{X}$, the optimal intervention characterizes optimal Y when intervening $\boldsymbol{X}$. This paper introduces the causal coverage indicator $c_{\boldsymbol{X}}$ with submodularity to approximately estimate $\mathbb{E}[Y|do(\boldsymbol{X}=\boldsymbol{x}^*)]$. Leveraging its submodular properties, it efficiently identifies a high-quality ES, cf. Sec. 4.1.2.

The causal coverage $c_{\boldsymbol{X}}$ can be calculated based on the topological structure of $\mathcal{G}$, with a minimal intervention cost, involving two parts. In part (i), $\mathbb{E}[Y|do(I=i^*)]$ of each single treatment variable $I \in \boldsymbol{I}$ is statistically estimated by BO under small intervention cost. And then, the intervention weight $0 \le w_V \le 1$ is normalized from $\mathbb{E}[Y|do(I=i^*)]/Co(I)$, where $Co(\cdot)$ means the cost function. The pseudocode of part (i) is displayed in Appendix. In part (ii), cf. Fig.1 (b), the causal coverage $c_{\boldsymbol{X}}$ of the intervention set $\boldsymbol{X}$ is calculated from w_V where V is the variable on the paths between $\boldsymbol{X}$ and Y. Alg. 1 displays the process of $c_{\boldsymbol{X}}$ calculation according to the topological order of $\mathcal{G}$, cf. Lines 2∼8. The topological order ensures that Alg. 1 calculates $c_{\boldsymbol{X}}$ completely and non-repetitively in the directed acyclic graph $\mathcal{G}$.

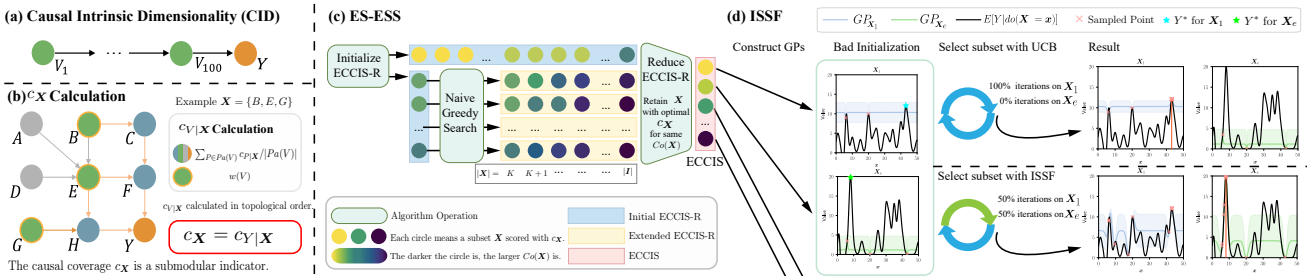

Figure 1. HCBO framework. (a) An example of CID $d_e = 1 \ll 100$. (b) A calculation example of the proposed submodular indicator $c_{\{B,E,G\}}$. (c) The proposed ES-ESS searches a polynomial approximation of MIS($\mathcal{M}$). (d) The proposed ISSF is capable of causal intrinsic dimensionality set $\boldsymbol{X}_e^*$ identification.

When intervening $\boldsymbol{X}$, the causal coverage of a single variable V is called conditional causal coverage, denoted as $c_{V|\boldsymbol{X}}$. In Line 1, $c_{V|\boldsymbol{X}}$ is initialized as 0, $\forall V \in \boldsymbol{V}$. Then, following the topological order, if $V \in \boldsymbol{X}$, then $c_{V|\boldsymbol{X}} = w_V$. Otherwise, $c_{V|\boldsymbol{X}}$ is the average of the conditional causal coverage of all of V's parent variables, denoted as $\sum_{P \in Pa(V)} c_{P|\boldsymbol{X}}/|Pa(V)|$. Here, the structural equation f_V is assumed to be additive and can be easily modified based on real-world priors. Finally, the causal coverage of $\boldsymbol{X}$ equals the conditional causal coverage of Y, i.e., $c_{\boldsymbol{X}} = c_{Y|\boldsymbol{X}}$.

4.1.2 *Efficient Causal Coverage Intervention Set (ECCIS)*

This section first analyzes and proves the submodularity of the proposed causal coverage indicator, cf. Thm. 1. Subsequently, leveraging the submodularity, ES-ESS determines the exploration set ECCIS($\mathcal{M}$) with polynomial sample complexity, cf. Alg. 2. If CID exists, the optimal intervention set satisfies $d_e = |\boldsymbol{X}_e^*| < |\boldsymbol{I}|$. ES-ESS is capable of finding a $(1 - 1/e)$-approximation of $c_{\boldsymbol{X}_e^*}$ of $\boldsymbol{X}_e^*$ within polynomial sample complexity. Moreover, Thm. 2 presents the existence theorem of CID, clearly defining the applicability scope of ES-ESS.

Theorem 1 (Submodularity of Causal Coverage Indicator)**.** *For all intervention set $\boldsymbol{X}_A \subseteq \boldsymbol{X}_B \subseteq \boldsymbol{I}$ and single treatment variable $I \in \boldsymbol{I} \setminus \boldsymbol{X}_B$, is hold that $c_{\boldsymbol{X}_A \cup \{I\}} - c_{\boldsymbol{X}_A} \geq c_{\boldsymbol{X}_B \cup \{I\}} - c_{\boldsymbol{X}_B}$.*

Theorem 2 (Causal Intrinsic Dimensionality Existence Theorem)**.** *If $Pa(Y) \subset \boldsymbol{V}$ and $\boldsymbol{I} = \boldsymbol{V}$, then the CID d_e exist, i.e., $d_e < |\boldsymbol{V}|$.*

Remark 1. *Thm. 1 and Thm. 2 are proved in Appendix. Thm. 1 demonstrates that the causal coverage $c_{\boldsymbol{X}}$ is a set indicator with submodularity, which manifests as diminishing marginal returns. Therefore, ES-ESS can identify ECCIS($\mathcal{M}$) with polynomial sample complexity via a $(1 - 1/e)$-competitive submodular optimization algorithm. The submodularity of causal coverage ensures the quality of intervention sets within ECCIS($\mathcal{M}$). Thm. 2 provides a sufficient but not necessary condition. Thm. 2 elucidates that a general CGO problem with $\boldsymbol{I} = \boldsymbol{V}$ generally has the CID, as $\mathcal{G}$ usually contains variables not connected to the target variable.*

ES-ESS outlines the process of calculating ECCIS($\mathcal{M}$) based on the submodular indicator $c_{\boldsymbol{X}}$, cf. Alg. 2. ES-ESS is inspired by a submodular optimization algorithm called Guess-K [19, 38, 13]. From the theorem of Guess-K, it is a $(1-1/e)$-competitive algorithm when the constant K = 3 [19]. Specifically, Guess-K first enumerates the intervention sets for $|\boldsymbol{X}| \leq$ K, cf. Line 1. Then, initialized with the intervention set of $|\boldsymbol{X}| =$ K, Guess-K uses greedy search for $|\boldsymbol{X}| >$ K, cf. Line 2∼9. ES-ESS diverges from classical submodular optimization algorithms by retaining intervention sets $\boldsymbol{X}$ with high causal coverage at different costs, cf. Line 10. ES-ESS preserves only the optimal intervention sets under the same cost to form ECCIS($\mathcal{M}$).

Algorithm 2 ES Effective Search Strategy (ES-ESS)

Input: Structural causal model $\mathcal{M}$, Cost function $Co(\cdot)$, Weights $\boldsymbol{w}$
Output: Exploration set ECCIS($\mathcal{M}$)

1: Initialize redundant ECCIS-R($\mathcal{M}$) $\leftarrow \{\boldsymbol{X} | \boldsymbol{X} \subseteq \boldsymbol{I}, |\boldsymbol{X}| \leq \text{K}\}$
2: **for** all $\boldsymbol{X} \subseteq \boldsymbol{I}$ with $|\boldsymbol{X}| =$ K **do**
3: $\quad \boldsymbol{W} \leftarrow \boldsymbol{I} \setminus \boldsymbol{X}$
4: $\quad$ **repeat**
5: $\quad\quad \boldsymbol{X} \leftarrow \boldsymbol{X} \cup \{W^*\}$, $W^* = \arg\max_{W \in \boldsymbol{W}} c_{\{\boldsymbol{X} \cup \{W\}\}}$, cf. Alg. 1
6: $\quad\quad$ ECCIS-R($\mathcal{M}$) $\leftarrow$ ECCIS-R($\mathcal{M}$) $\cup \boldsymbol{X}$
7: $\quad\quad \boldsymbol{W} \leftarrow \boldsymbol{W} \setminus W^*$
8: $\quad$ **until** $\boldsymbol{W} = \emptyset$
9: **end for**
10: ECCIS($\mathcal{M}$) $\leftarrow$ ECCIS-R($\mathcal{M}$) after removing $\boldsymbol{X}$ that is suboptimal for causal coverage $c_{\boldsymbol{X}}$ at the same cost $Co(\boldsymbol{X})$
11: **return** ECCIS($\mathcal{M}$)

4.2 *Intervention Set Score Function (ISSF)*

The standard CBO framework constructs the GP models on the intervention dataset, denoted as $GP_{D^I_{\boldsymbol{X}}}$, $\forall \boldsymbol{X} \in$ ES. The acquisition function is utilized to identify the causal intrinsic intervention set $\boldsymbol{X}_e^*$. Since the actual $\boldsymbol{X}_e^*$ cannot be determined in the algorithm process, this paper denotes the perceived one during algorithm process as $\hat{\boldsymbol{X}}_e^*$. Then, the same acquisition function is used to determine the best intervention $\boldsymbol{x}^*$ within $D(\hat{\boldsymbol{X}}_e^*)$. However, general acquisition functions are not suitable for the comparison of sampling points between multiple GP models, because of the scale inconsistency of target variable Y among them. This inconsistency manifests as significant differences in $D(Y)$ across GP models. In the high-dimensional CGO problem, increasing the size of ES aggravates the negative impact of scale inconsistency.

The scale inconsistency can bias the GP model selection process, potentially overlooking $\boldsymbol{X}_e^*$ due to bad initialization, leading to local optima. Common GP acquisition functions like expected improvement [18] (EI) and upper confidence bound [36] (UCB) are designed to evaluate sampling points within the same GP and are not suited for selecting the optimal GP model. To overcome this issue, this paper employs an intervention set score function $ISSF(\boldsymbol{X})$, cf. Eq. (2). This scoring function is essentially a class of normalized UCB, which uses the normalized intervention dataset $\tilde{D}^I_{\boldsymbol{X}}$. The target variable of $\tilde{D}^I_{\boldsymbol{X}}$ follows zero mean and unit variance. $ISSF(\boldsymbol{X})$ effectively mitigates the scale inconsistency between GP models, thus optimizing the overall model selection process and preventing local optima.

$$ISSF(\boldsymbol{X}) = \bar{Y}_{D^I_{\boldsymbol{X}}} + \beta_1 UCB(\boldsymbol{x}^*_{\boldsymbol{X}} | GP_{\tilde{D}^I_{\boldsymbol{X}}}) . \quad (2)$$

$\bar{Y}_{D^I_{\boldsymbol{X}}}$ means the mean of target variable in the intervention dataset. $UCB(\boldsymbol{x}^*_{\boldsymbol{X}} | GP_{\tilde{D}^I_{\boldsymbol{X}}})$ means the optimal UCB of $GP_{\tilde{D}^I_{\boldsymbol{X}}}$ in the normalized intervention dataset, where $UCB(\boldsymbol{x}^* | GP_{\tilde{D}^I_{\boldsymbol{X}}}) = \mu(\boldsymbol{x}^*) + \beta_2 \sigma(\boldsymbol{x}^*)$. $\mu(\boldsymbol{x}^*)$ and $\sigma(\boldsymbol{x}^*)$ are the mean and variance predictions of Y at sampling point $\boldsymbol{x}^*$. $\boldsymbol{x}^*$ means the optimal sampling point found by $GP_{\tilde{D}^I_{\boldsymbol{X}}}$ with UCB acquisition function. Meanwhile, β_1, β_2 are appropriate constants and specified according to the context. Specifically, β_1 aligns the scale of $\bar{Y}_{D^I_{\boldsymbol{X}}}$ with $UCB(\boldsymbol{x}^*_{\boldsymbol{X}} | GP_{\tilde{D}^I_{\boldsymbol{X}}})$, while β_2 is derived from the theorem of UCB [36]. Constants β_1, β_2 are set to balance exploration and exploitation [42].

4.3 The HCBO Algorithm

This paper introduces the HCBO algorithm, capable of solving high-dimensional CGO problems, cf. Alg. 3. The proposed ES-ESS finds ECCIS($\mathcal{M}$), which is a polynomial-scale approximation of the precise MIS($\mathcal{M}$), cf. Line 2. Moreover, the proposed ISSF overcomes the scale inconsistency issue between GP model comparison, ensuring the identification of $\boldsymbol{X}^*_e$ within ECCIS($\mathcal{M}$), cf. Line 5.

Algorithm 3 HCBO

Input: Structural causal model $\mathcal{M}$, Intervention dataset D^I
Output: Causal intrinsic intervention set $\boldsymbol{X}^*_e$, Optimal target y^*

1: $\tilde{D}^I \leftarrow$ normalizing $(\boldsymbol{x}, y) \in D^I$ to $(\boldsymbol{x}, \frac{y - \mu(D^I)}{\sigma(D^I)})$
2: Calculate the exploration set ECCIS($\mathcal{M}$) with Alg. 2
3: Initialize GP model $GP_{\tilde{D}^I_{\boldsymbol{X}}}, \forall \boldsymbol{X} \in$ ECCIS($\mathcal{M}$)% $[T]$ means a monotonically increasing ordered set consisting of $1, \ldots, T$.
4: **for** $t \in [T]$ **do**
5: Select $\hat{\boldsymbol{X}}^*_{e,t} \in$ ECCIS($\mathcal{M}$) via $ISSF(\boldsymbol{X})$
6: Select $\boldsymbol{x}^*_t \in D(\hat{\boldsymbol{X}}^*_{e,t})$ via UCB acquisition function
7: Intervene $\mathcal{M}$ to get $y_t = \mathbb{E}[Y | do(\hat{\boldsymbol{X}}^*_{e,t} = \boldsymbol{x}^*_t)]$
8: Update datasets $D^I_{\boldsymbol{X}_t}$ and $\tilde{D}^I_{\boldsymbol{X}_t}$ via $(\boldsymbol{x}^*_t, y_t)$
9: Update the surrogate model $GP_{\tilde{D}^I_{\boldsymbol{X}^*_t}}$
10: **end for**
11: **return** $\hat{\boldsymbol{X}}^*_{e,T}, y^* = \mathbb{E}[Y | do(\hat{\boldsymbol{X}}^*_{e,T} = \boldsymbol{x}^*_T)]$

5 Experiments

Experiments are conducted to answer the following research questions. **Q1:** Is causal intrinsic dimensionality (CID) widely present in CGO problems? **Q2:** How do HCBO's scalability, convergence speed, and final optimum compare to existing methods across various CGO problems? **Q3:** How much better can the exploration set (ES) found by ES-ESS component be than the usual strategy? **Q4:** Why is ISSF component better at selecting $\hat{\boldsymbol{X}}_e$ than others? **Q5:** How to automate the selection of optimal hyperparameter β_1?

5.1 Experiment Setup

Concept Definition. The structural equations $f_v \in \boldsymbol{F}$ of $\mathcal{M}$ here are categorized into three types. (a) Linear: $f_V(Pa(V), U_V) = \sum_{P \in Pa(V)} w_{VP} x_P + b_V + U_V$, where w_{VP} and b_V are coefficients; (b) Additive: $f_V(Pa(V), U_V) = \sum_{P \in Pa(V)} w_{VP} h_{VP}(x_P) + b_V + U_V$, where h_{VP} is identity transformation or non-linear transformation like log; (c) Non-Additive: $f_V(Pa(V), U_V) = h_V(Pa(V)) + U_V$, where h_V can be complex functions.

Datasets. Experiments are conducted on varieties of synthetic and real-world settings. High-dimensional CGO problems are generated by a random generator, inspired by the gCastle toolkit [43]. They include linear-100-33, linear-200-66, additive-50-16, additive-100-33, non-additive-50-16, non-additive-100-33. Consider the example "linear-100-33". The name indicates that this problem follows linear structural equations and includes 100 endogenous observable variables, 33 of which are treatment variables. All of them aim to maximize the variable positioned at the end of its topological order, denoted as Y. Meantime, real-world CGO tasks include Coral Ecology [9] and Health [14]. Dataset Coral Ecology considers various factors of a coral ecosystem with 11 endogenous observable variables, 5 of which are treatment variables. The SCM follows linear structural equation, and aims to maximize net coral ecosystem calcification [1]. Dataset Health is based on the SCM built by [14]. It models the causal effect of statin drugs on the levels of prostate specific antigen (PSA). It includes 9 endogenous observable variables, 3 of which are treatment variables. Health follows non-additive structural equations and aims to minimize PSA. Details are explained in Appendix.

Compared Methods. To the best of our knowledge, current research lacks methods that effectively use causation for high-dimensional CGO problems like HCBO. For a comprehensive analysis of HCBO's superiority, we compare it with various methods, including engineering optimization methods, standard CBO method, and HBO methods. Especially, **Random Search**: A method that randomly selects a subset $\boldsymbol{X} \subseteq \boldsymbol{I}$, and randomly selects intervention values $\boldsymbol{x} \in D(\boldsymbol{X})$; **BO**: A standard Bayesian optimization method; **CMA-ES** [17]: A widely used evolutionary method for black-box optimization; **CBO** [1]: A classical method utilizes causation by constructing multiple GPs for optimization; **REMBO** [41]: A pioneering study that utilizes random linear embedding for projecting high dimensions to lower dimensions; **ALEBO** [21]: An embedding BO method enhanced by Mahalanobis kernel and hypersphere sampling; **Dropout-BO** [22]: A variable selection BO method that chooses d dimensions out of D dimensions randomly and optimizes variables from the chosen dimensions via BO; **MCTS-VS** [35]: A variable selection BO method that uses Monte-Carlo tree to search d dimensions out of D dimensions to optimize; **TuRBO** [12]: A BO method that fits a collection of local models to optimize in the trust region. Please refer to the Appendix for detailed settings of these methods.

Standard optimization methods are not designed to solve CGO problems. In the compared methods above, standard optimization methods are BO, CMA-ES, REMBO, ALEBO, Dropout-BO, MCTS-VS and TuRBO. We reformulate CGO problem to standard optimization problem as Eq. (3). This reformulation makes these standard optimization methods intervene on all treatment variables to optimize.

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x} \in D(\boldsymbol{I})} \mathbb{E}\left[Y | do(\boldsymbol{I} = \boldsymbol{x})\right] . \quad (3)$$

Experimental Settings. Baseline experiments are repeated independently for 20 times. Intervention cost of all treatment variables is set as 1. All baselines share the same cost budget. For a fair comparison, in all experiments except for ablation studies, HCBO's intervention cost spent in calculating the intervention weight is considered. For clearer presentation, all outcomes Y of synthetic problems are normalized to the range $[0, 1]$ before display, cf. Fig.2. The full experimental details can be found in Appendix.

5.2 CID Validation Experiment (To Q1)

The validation experiments are conducted to show that CID is widely seen in CGO problems. Specifically, we sample intervention sets and

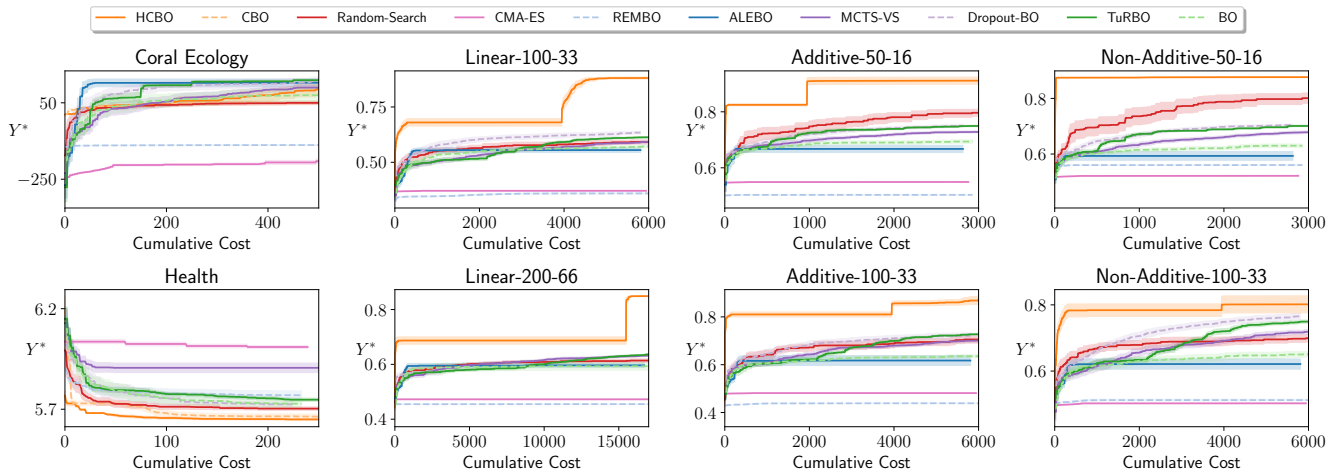

Figure 2. Optimization performance comparison on 8 representative CGO problems. HCBO demonstrates SOTA convergence speed, target values Y^*.

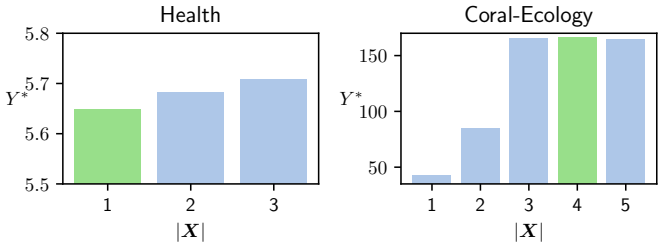

Figure 3. CID validation in real-world tasks. The green bar means the causal intrinsic intervention set $\boldsymbol{X}_e^*$.

obtain the corresponding optimal interventions by TuRBO to approximately determine $\boldsymbol{X}_e^*$. To sample intervention sets, we enumerate all subsets in real-world tasks, while using Monte-Carlo method like [45] do in high-dimensional settings, due to the curse of dimensionality. The existence of CID is validated across all problems involved in this paper. Fig.3 displays the CID of the real-world tasks, i.e., Coral Ecology and Health. Other results and details are displayed in Appendix. Fig.3 reveals that intervening all variables $\boldsymbol{I}$ in CGO problems may result in local optima, restrengthening importance to utilize causation for reaching the global optimum.

5.3 *Performance Experiment (To Q2)*

The performance experiment is conducted for comparing HCBO against all compared methods across synthetic problems and real-world tasks. We also perform t-test to compare final optimums achieved by HCBO and other methods, which indicates that HCBO outperforms all compared methods in all datasets except Coral Ecology. Due to page limitation, the statistical results and more details can be found from Appendix.

5.3.1 *Optimization of Synthetic Benchmark*

In terms of the number of treatment variables, our synthetic high–dimensional CGO problems extend previous explorations within this field. The synthetic problems vary in the numbers of variables and structural equations. We show the results in Fig.2, illustrating the following observations: (a). HCBO achieves SOTA performance in all high-dimensional problems. It outperforms random search, a robust baseline in high-dimensional CGO scenarios, demonstrating the superiority of HCBO in searching for optimal intervention sets from an exponential search space. It defeats standard high dimensional methods because it can utilize causation, while others can't. In addition, HCBO demonstrates effective initialization and converge quickly. The former is attributed to the proposed intervention weight estimation, and the latter is attributed to the ability for identifying CID. (b). HCBO demonstrates significant advantages in high-dimensional complex problems involving the non-additive structural equations. Specifically, many standard methods can optimize low-dimensional non-additive problem Health, but they fail in the high-dimensional complex problems non-additive-100-124. The reason are twofold: HBO methods fail to detect the effective dimensionality since it may not exist in CGO problems, and HCBO can locate CID and the correspond optimal intervention set.

5.3.2 *Optimization of Real-World Tasks*

On Coral Ecology Dataset. As illustrated in Fig.2 and statistical results (in Appendix), HCBO does not outperform all compared methods. This can be explained with Fig.3, which shows the optimal target value of $\boldsymbol{I}$ is similar to that of $\boldsymbol{X}_e^*$. The similarity reduces the importance of selecting $\boldsymbol{X}_e^*$ but emphasizes the need to fully exploit the potential of $\boldsymbol{X}_e^*$. Since HCBO is not good at optimizing a particular subset, HCBO converges slower than some compared methods. This indicates room for further improvement of HCBO.

On Health Dataset. In the harder non-additive real-world with CID $d_e = 1$ as shown in Fig.3, standard optimization methods, including ALEBO that converges prematurely, fail to reach global optimum, but HCBO and CBO can reach, cf. Fig.2. In addition, HCBO defeats CBO in two respects. Firstly, HCBO identifies $\boldsymbol{X}_e^*$ through ES-ESS, ensuring that there is chance for HCBO to achieve global optimum. Secondly, the found ECCIS($\mathcal{M}$) with size 3 is smaller than CBO's MIS($\mathcal{M}$) with size 7. Therefore, HCBO can save cost on unnecessary intervention sets and pay more cost to fully search and optimize $\boldsymbol{X}_e^*$.

5.4 *Ablation Study: Exploration Set (To Q3)*

We compare ECCIS with two types of randomly initialized exploration sets, including (a) **SCRIS** Same cost random intervention set; (b) **DCRIS** Different cost random intervention set. For a fair comparison, they both share the same number of intervention sets as ECCIS. Specifically, SCRIS must cover all different costs in ECCIS. DCRIS are randomly initialized without this constraint. Results on the representative high-dimensional problem additive-100-33 in Fig.4 show that ECCIS significantly outperforms others. Even the best result among all random exploration sets are inferior to that of ECCIS.

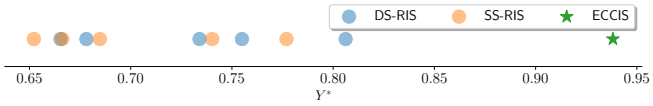

Figure 4. Comparison between ECCIS and other exploration sets (repeated 5 times). Each scatter means the optimum during optimization.

5.5 *Ablation Study: ISSF (To Q4)*

To explain the advantages of ISSF, ISSF is compared with varieties of settings to demonstrate its ability to stably identify CID, cf. Fig.5 and Fig.6. We compare varieties of CID identification methods, including UCB, EI, Mean and ISSF. CID identification methods score each intervention set in ECCIS$(\mathcal{M})$, and select the intervention set with the highest score to intervene. General acquisition functions, like UCB and EI, score intervention sets with their optimal acquisition function values. Mean method utilizes the mean of target variable in $D^I_{\boldsymbol{X}}$ to score $\boldsymbol{X}$. Proposed HCBO utilizes ISSF to score $\boldsymbol{X}$. Since HCBO

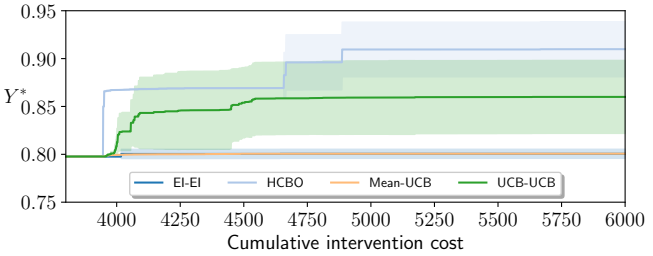

Figure 5. Comparison of set selection functions: ISSF, UCB, Mean, and EI.

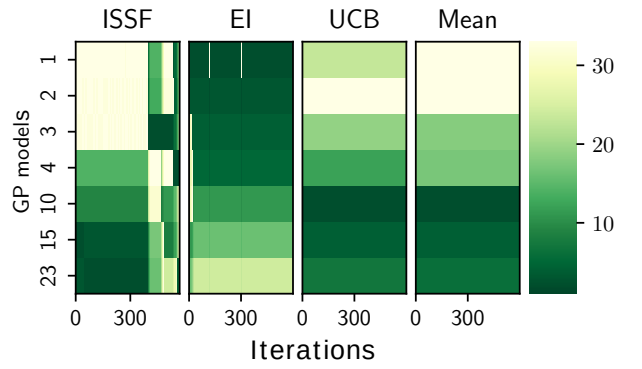

Figure 6. The update of estimated ranks of multiple GP models by different set selection functions during 600 optimization iterations. The darker the GP model is, the higher the rank is. For each iteration, HCBO selects the intervention set ranking the first to intervene. The horizontal color variation indicates the update of the global information with optimization iteration.

uses UCB acquisition function to select optimal intervention value $\boldsymbol{x}^*$, HCBO is denoted as ISSF-UCB here. We compare the following methods with our ISSF-UCB: UCB-UCB, Mean-UCB, EI-EI. We compare EI-EI as it is used by CBO [1]. The overall performance comparison in Fig.5 demonstrates that HCBO outperforms all other variants. For other variants, UCB-UCB outperforms Mean-UCB and EI-EI, for Mean-UCB and EI-EI fail to optimize beyond the results of the initial dataset. We research why this happens by conducting a case study as shown in Fig.6, and we note the following observations. (a) Compared to others, ISSF evaluates different intervention sets to update global information. In Fig.6, HCBO exploits GP model $15, 23, 3, \ldots$ respectively. This helps balance exploration and exploitation on intervention sets. (b) EI fails to estimate the potential of intervention sets, especially in high-dimensional scenarios. Its scores on intervention sets decrease according to the size of the intervention sets. This feature makes it only focus on small intervention sets, ignoring the potential of large intervention sets. (c) UCB and Mean can detect good high dimensional intervention sets. However, they fail to adjust their preferences for intervention. They perform over-exploitation on a single intervention set, falling into local optima.

5.6 *Hyperparameter Analysis (To Q5)*

We compare different strategies to select the hyperparameter β_1 within the ISSF component, including static strategies, where β_1 is fixed at a specific value, and dynamic strategies, where β_1 is updated periodically. We find that: (a). Dynamic strategies prevail static strategies. (b). Dynamic strategy performs better when β_1 is updated more frequently. We choose a conservative dynamic approach "Average-50" for HCBO. Further details are provided in Appendix.

6 Conclusion

This study introduces high-dimensional causal Bayesian optimization (HCBO) as an innovative solution to CBO challenges, utilizing causal intrinsic dimensionality (CID) to address the curse of dimensionality from the number of variables of the causal graph and scale inconsistency among GP models. HCBO represents a significant advancement, capable of managing causal graphs up to 200 sizes, markedly improving from the previous limit of about 10. HCBO's contribution lies in proposing an efficient search strategy for CID, i.e., ES-ESS, which makes CID theory practically applicable. Additionally, HCBO addresses the issue of scale inconsistency when comparing multiple GP models, which can invalidate general acquisition functions, and offers a simple and feasible solution, i.e., ISSF.

Despite progress, HCBO's effectiveness is more suitable for SCM following additive equation assumption. Designing a targeted calculation method of causal coverage for non-additive causal graphs would be addressed in future work. In addition, future directions would focus on enhancing autonomous generation and validation of causal graphs, as well as developing adaptive scaling within the ISSF to broaden HCBO's applicability across diverse high-dimensional problems.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive and helpful reviews and suggestions. This work is supported by the National Natural Science Foundation of China (No. 62106076) and Ant Research Program.

References

[1] V. Aglietti, X. Lu, A. Paleyes, and J. González. Causal Bayesian optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3155–3164, Sicily, Italy, 2020.

[2] V. Aglietti, N. Dhir, J. González, and T. Damoulas. Dynamic causal Bayesian optimization. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 10549–10560, virtual, 2021.

[3] V. Aglietti, A. Malek, I. Ktena, and S. Chiappa. Constrained causal Bayesian optimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 304–321, Honolulu, HI, 2023.

[4] K. Antonov, E. Raponi, H. Wang, and C. Doerr. High dimensional Bayesian optimization with kernel principal component analysis. In *Proceedings of the 17th International Conference on Parallel Problem Solving from Nature (PPSN)*, volume 13398, pages 118–131, Dortmund, Germany, 2022.

[5] M. Binois, D. Ginsbourger, and O. Roustant. A warped kernel improving robustness in Bayesian optimization via random embeddings. In *Proceedings of the 9th International Conference on Learning and Intelligent Optimization (LION)*, volume 8994, pages 281–286, Lille, France, 2015.

[6] M. Binois, D. Ginsbourger, and O. Roustant. On the choice of the low-dimensional domain for global optimization via random embeddings. *Journal of Global Optimization*, 76(1):69–90, 2020.

[7] N. Branchini, V. Aglietti, N. Dhir, and T. Damoulas. Causal entropy optimization. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 8586–8605, Valencia, Spain, 2023.

[8] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. SIAM, Philadelphia, PA, 2009.

[9] T. A. Courtney, M. Lebrato, N. R. Bates, A. Collins, S. J. De Putron, R. Garley, R. Johnson, J.-C. Molinero, T. J. Noyes, C. L. Sabine, et al. Environmental controls on modern scleractinian coral and reef-scale calcification. *Science advances*, 3(11):e1701356, 2017.

[10] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Finite sample convergence rates of zero-order stochastic optimization methods. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, pages 1448–1456, Lake Tahoe, NV, 2012.

[11] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[12] D. Eriksson, M. Pearce, J. R. Gardner, R. Turner, and M. Poloczek. Scalable global optimization via local Bayesian optimization. In *Advances in Neural Information Processing Systems 32*, pages 5497–5508, Vancouver, Canada, 2019.

[13] M. Feldman, Z. Nutov, and E. Shoham. Practical budgeted submodular maximization. *Algorithmica*, 85(5):1332–1371, 2023.

[14] A. Ferro, F. Pina, M. Severo, P. Dias, F. Botelho, and N. Lunet. Use of statins and serum levels of prostate specific antigen. *Acta Urológica Portuguesa*, 32(2):71–77, 2015.

[15] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[16] L. Gultchin, V. Aglietti, A. Bellot, and S. Chiappa. Functional causal Bayesian optimization. In R. J. Evans and I. Shpitser, editors, *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence UAI*, volume 216, pages 756–765, Pittsburgh, PA, 2023.

[17] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, 2003.

[18] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.*, 13(4):455–492, 1998.

[19] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Inf. Process. Lett.*, 70(1):39–45, 1999.

[20] S. Lee and E. Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 2573–2583, Montréal, Canada, 2018.

[21] B. Letham, R. Calandra, A. Rai, and E. Bakshy. Re-examining linear embeddings for high-dimensional Bayesian optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 1546–1558, virtual, 2020.

[22] C. Li, S. Gupta, S. Rana, V. Nguyen, S. Venkatesh, and A. Shilton. High dimensional Bayesian optimization using dropout. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2096–2102, Melbourne, Australia, 2017.

[23] Y. Liu, Q. Du, Q. Wang, H. Yu, J. Liu, Y. Tian, C. Chang, and J. Lei. Causal inference between bioavailability of heavy metals and environmental factors in a large-scale region. *Environmental pollution*, 226: 370–378, 2017.

[24] Y. Liu, Y. Hu, H. Qian, Y. Yu, and C. Qian. ZOOpt: A toolbox for derivative-free optimization. *Science China Information Sciences*, 65: 207101, 2022.

[25] A. Nayebi, A. Munteanu, and M. Poloczek. A framework for Bayesian optimization in embedded subspaces. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4752–4761, Long Beach, CA, 2019.

[26] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688, 1995.

[27] J. Pearl. *Causality*. Cambridge University Press, 2009.

[28] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

[29] H. Qian and Y. Yu. Derivative-free reinforcement learning: A review. *Frontiers of Computer Science*, 15(6):156336, 2021.

[30] H. Qian, Y. Hu, and Y. Yu. Derivative-free optimization of high-dimensional non-convex functions by sequential random embeddings. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1946–1952, New York, NY, 2016.

[31] L. M. Rios and N. V. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *J. Glob. Optim.*, 56(3):1247–1293, 2013.

[32] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[33] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, and A. G. Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021.

[34] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, pages 2960–2968, Lake Tahoe, NV, 2012.

[35] L. Song, K. Xue, X. Huang, and C. Qian. Monte Carlo tree search based variable selection for high-dimensional Bayesian optimization. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 28488–28501, New Orleans, LA, 2022.

[36] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1015–1022, Haifa, Israel, 2010.

[37] S. Sussex, A. Makarova, and A. Krause. Model-based causal Bayesian optimization. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.

[38] M. Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Oper. Res. Lett.*, 32(1):41–43, 2004.

[39] C. Thompson. Causal graph analysis with the causalgraph procedure. In *Proceedings of SAS Global Forum*, Dallas, TX, 2019.

[40] A. Tripp, E. A. Daxberger, and J. M. Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 11259–11272, virtual, 2020.

[41] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. de Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

[42] H. Wattez, F. Koriche, C. Lecoutre, A. Paparrizou, and S. Tabary. Learning variable ordering heuristics with multi-armed bandits and restarts. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, volume 325, pages 371–378, Santiago de Compostela, Spain, 2020.

[43] K. Zhang, S. Zhu, M. Kalander, I. Ng, J. Ye, Z. Chen, and L. Pan. gCastle: A python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*, 2021.

[44] M. Zhang, H. Li, and S. W. Su. High dimensional Bayesian optimization via supervised dimension reduction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4292–4298, Macao, China, 2019.

[45] Y. Zhang, H. Qian, X. Shu, and A. Zhou. High-dimensional dueling optimization with preference embedding. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence, February*, pages 11280–11288, Washington, DC, 2023.