Appendix of "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]East China Normal University, China

The appendix contains the details of the method, proof, and experiment. Specifically, the method details show how the proposed HCBO calculates the intervention weights, cf. Alg. 1.The proof details demonstrate the submodularity property of our proposed causal coverage indicator (cf. Thm. 1) and the existence of the causal intrinsic dimensionality (CID) (cf. Thm. 2). Finally, the experiment details introduce the dataset, the specific implementation of the method, and additional experiments.

A Notation

The definitions of variables and parameters involved in this paper are listed in Table 1.

B Method Details

We estimate $\mathbb{E}[Y|do(I = i^*)]$ for each single treatment variable $I \in \boldsymbol{I}$ with Alg.1. $\mathbb{E}[Y|do(I = i^*)]$ is noted as intervention weight w_v. For each V, we calculate w_V by solving a single variable optimization problem. We utilize Bayesian Optimization technique with UCB acquisition function to handle it. The estimation process requires a small intervention cost, cf. Line 6.

Algorithm 1 Endogenous Observable Variable Weight Analysis

Input: Structural Causal Model $\mathcal{M}$, Cost function $Co(\cdot)$
Output: Weights $\boldsymbol{w}$ of endogenous observable variables

1: **for** $V \in \boldsymbol{V}$ **do**
2: $D_V^{I'} \leftarrow$ Collect intervention data of variable V from $\mathcal{M}$
3: Initialize the surrogate model $GP_{D_V^{I'}}$
4: **for** $t \in [T]$ **do**
5: $v_t^* \leftarrow$ Optimize $GP_{D_V^{I'}}$ via UCB acquisition function
6: Intervene $\mathcal{M}$ to get $y_t = \mathbb{E}[Y|do(V = v_t^*)]$
7: Update $D_V^{I'} \leftarrow D_V^{I'} \cup (v_t^*, y_t)$
8: **end for**
9: $w_V \leftarrow \max_{t\in[T]} y_t / Co(V)$
10: **end for**
11: **return** $\boldsymbol{w}$ after normalized by min-max scaling

* Corresponding Author. Email: hqian@cs.ecnu.edu.cn.
[1] Equal contribution.

C Proof Details

Algorithm 2 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** $c_{\boldsymbol{X}} \leftarrow c_{Y|\boldsymbol{X}}$

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

Proof. The calculation of the causal coverage indicator is outlined in the algorithm: Calculation of Causal Coverage of $\boldsymbol{X}$, which is donated as Alg. 2 in Sec.4 of the main paper. It is shown in the appendix as Alg. 2.

In the process of proving $c_{\boldsymbol{X}_A\cup I} - c_{\boldsymbol{X}_A} \geq c_{\boldsymbol{X}_B\cup I} - c_{\boldsymbol{X}_B}$, the complex calculation of causal coverage is broken down into the calculation of conditional causal coverage.

Firstly, the characteristics of the causal coverage indicator are introduced: (a). $c_\emptyset = 0$, i.e., the causal coverage indicator is 0 when no treatment variables are intervened. (b). The conditional causal coverage of treatment variables that are not on the connected path from the intervention set $\boldsymbol{X}$ to Y (denoted as $Path_{\boldsymbol{X}}$) is 0. (c). $c_{V|\boldsymbol{X}}, \forall V \in \boldsymbol{V}$, i.e., the conditional causal coverage of every treatment variable is 0 due to the intervention weight $w_V \geq 0$.

Thereafter, the proof begins: Intervening $\boldsymbol{X}$ alters the structure of the causal graph $\mathcal{G}$, severing all causal connections entering the intervention node set. Variable $I \in \boldsymbol{I} \setminus \boldsymbol{X}$ can be categorized into three: (a). the variable I satisfies both I is independent of Y given $do(\boldsymbol{X}_A)$ (denoted as $I \perp\!\!\!\perp Y|do(\boldsymbol{X}_A)$) and I is independent of Y given $do(\boldsymbol{X}_B)$ (denoted as $I \perp\!\!\!\perp Y|do(\boldsymbol{X}_B)$), (b). the variable I satisfies both $I \not\perp\!\!\!\perp Y|do(\boldsymbol{X}_A)$ and $I \not\perp\!\!\!\perp Y|do(\boldsymbol{X}_B)$, and (c). the variable I satisfies both $I \not\perp\!\!\!\perp Y|do(\boldsymbol{X}_A)$ and $I \perp\!\!\!\perp Y|do(\boldsymbol{X}_B)$. Note that no variable I exists that satisfies both $I \perp\!\!\!\perp Y|do(\boldsymbol{X}_A)$ and $I \not\perp\!\!\!\perp Y|do(\boldsymbol{X}_B)$ because $\boldsymbol{X}_A \subseteq \boldsymbol{X}_B$.

Table 1. The notions involved in this paper.

Notation	Definition
$\mathcal{G}$	Directed acyclic causal graph.
$\boldsymbol{U}$	Set of exogenous variables.
$\boldsymbol{V}$	Set of endogenous variables.
$\boldsymbol{F}$	Set of functions which defines the functional relations between the variables in $\mathcal{G}$.
$P(\boldsymbol{U})$	Probability distribution of $\boldsymbol{U}$.
$\mathcal{M}$	Structural causal model, which contains $\mathcal{G}$ and four-tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$.
$\boldsymbol{C}$	Set of non-manipulative variables in $\boldsymbol{V}$.
$\boldsymbol{I}, I$	Set of treatment variables in $\boldsymbol{V}$ and a single treatment variable.
Y	Target variable in $\boldsymbol{V}$.
$Pa(V)$	Parent variables of a single variable V.
$\mathcal{P}(\cdot)$	Power set.
ES	Exploration set, which can be searched by $\mathcal{P}(\boldsymbol{I})$, MIS($\mathcal{M}$) and ECCIS($\mathcal{M}$).
$\boldsymbol{X}$	Intervention set and $\boldsymbol{X} \in \boldsymbol{I}$.
$D(\boldsymbol{X})$	$\bigtimes_{I \in \boldsymbol{X}} D(I)$, where $D(I)$ is intervention domain of I.
$Co(\boldsymbol{X})$	Intervention cost of $\boldsymbol{X}$.
$do(\boldsymbol{X} = \boldsymbol{x})$	Interventional operation on $\boldsymbol{X}$ with fixing its value to $\boldsymbol{x}$.
$P(Y \vert do(\boldsymbol{X} = \boldsymbol{x}))$	Probability distribution of Y obtained by intervening on $\boldsymbol{X}$ with fixing its value to x.
$\mathbb{E}[Y \vert do(\boldsymbol{X} = \boldsymbol{x})]$	Interventional expectation corresponding to $P(Y \vert do(\boldsymbol{X} = \boldsymbol{x}))$.
$\mathbb{E}[Y \vert do(\boldsymbol{X} = \boldsymbol{x}^*)]$	Optimal intervention, where $\boldsymbol{x}^*$ is the optimal intervention value in $D(\boldsymbol{X})$.
$\boldsymbol{X}_e, \mathbb{X}_e$	Effective intervention set and the set of effective intervention set.
$\boldsymbol{X}_e^\star$	Causal intrinsic intervention, where $\boldsymbol{X}_e^* = \arg\min_{\boldsymbol{X}_e \in \mathbb{X}_e} Co(\boldsymbol{X}_e)$.
d_e	Causal intrinsic dimensionality, $d_e = \vert \boldsymbol{X}_e^\star \vert$.
$\hat{\boldsymbol{X}}_e^\star$	Causal intrinsic dimensionality found in the algorithm process.
w_V	Intervention weight of V.
$c_{\boldsymbol{X}}$	Causal coverage of $\boldsymbol{X}$.
$c_{V \vert \boldsymbol{X}}$	Conditional causal coverage of V when intervening $\boldsymbol{X}$.
$D^I, D^I_{\boldsymbol{X}}$	Intervention dataset and sub-dataset gotten from intervening $\boldsymbol{X}$.
$\tilde{D}^I_{\boldsymbol{X}_t}$	Normalized intervention dataset. The target variable of $\tilde{D}^I_{\boldsymbol{X}_t}$ follows zero mean and unit variance.
$\bar{Y}_{D^I_{\boldsymbol{X}}}$	Mean of target variable in $D^I_{\boldsymbol{X}}$.
$UCB(\boldsymbol{x}^*_{\boldsymbol{X}} \vert GP_{\tilde{D}^I_{\boldsymbol{X}}})$	Optimal UCB of GP model building on $\tilde{D}^I_{\boldsymbol{X}}$.
β_1, β_2	Appropriate constants for balancing exploration and exploitation.
$Path_{\boldsymbol{X}}$	Set of connected path between X and Y.

(a). If I satisfies both $I \perp\!\!\!\perp Y|do(\boldsymbol{X}_A)$ and $I \perp\!\!\!\perp Y|do(\boldsymbol{X}_B)$, it indicates that at least one treatment variable in the intervention set $\boldsymbol{X}$ exists in every connected path from I to Y. The conditional causal coverage of intervention variable equals the intervention weight, cf. Alg. 2 Line 4 . Therefore, adding I to $\boldsymbol{X}_A$ or $\boldsymbol{X}_B$ yields no causal coverage gain, i.e., $c_{\boldsymbol{X}_A \cup I} - c_{\boldsymbol{X}_A} = c_{\boldsymbol{X}_B \cup I} - c_{\boldsymbol{X}_B} = 0$.

(b). If I satisfies both $I \not\perp\!\!\!\perp Y|do(\boldsymbol{X}_A)$ and $I \not\perp\!\!\!\perp Y|do(\boldsymbol{X}_B)$, it indicates that at least one connected path exists that does not include either treatment variable in $\boldsymbol{X}_A$ or $\boldsymbol{X}_B$, and the set of connected paths that do not encompass $\boldsymbol{X}_A$ includes the set of connected paths that do not encompass $\boldsymbol{X}_B$, denoted as $Path_{\boldsymbol{X}_B} \subseteq Path_{\boldsymbol{X}_A}$, $|Path_{\boldsymbol{X}_A}| \geq |Path_{\boldsymbol{X}_B}| \geq 1$. The conditional causal coverage of the non-intervention variable equals the arithmetic mean of the conditional causal intervention of its parent variables, cf. Alg. 2 Line 6. Since conditional causal coverage is not less than 0, the causal coverage gain brought by the connected path is also not less than 0. Since $Path_{\boldsymbol{X}_B} \subseteq Path_{\boldsymbol{X}_A}$, $c_{\boldsymbol{X}_A \cup I} - c_{\boldsymbol{X}_A} \geq c_{\boldsymbol{X}_B \cup I} - c_{\boldsymbol{X}_B} \geq 0$.

(c). If I satisfies both $I \not\perp\!\!\!\perp Y|do(\boldsymbol{X}_A)$ and $I \perp\!\!\!\perp Y|do(\boldsymbol{X}_B)$, it indicates that at least one treatment variable in $\boldsymbol{X}_B$ exists in every connected path from I to Y, but there is at least one connected path that does not include any treatment variable in $\boldsymbol{X}_A$. It is easy to know from (a) and (b) situations that $c_{\boldsymbol{X}_A \cup I} - c_{\boldsymbol{X}_A} \geq c_{\boldsymbol{X}_B \cup I} - c_{\boldsymbol{X}_B} = 0$.

In summary, the causal coverage indicator satisfies submodularity in all three situations, i.e., $c_{\boldsymbol{X}_A \cup I} - c_{\boldsymbol{X}_A} \geq c_{\boldsymbol{X}_B \cup I} - c_{\boldsymbol{X}_B}$. The proof is done. □

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* $\boldsymbol{X}_e^* = \arg\min_{\boldsymbol{X}_e \in \mathbb{X}_e} Co(\boldsymbol{X}_e)$.

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.,* $d_e < |\boldsymbol{V}|$.

Proof. It is obvious that $d < D$, then for the sake of contradiction, we assume $d = D$, and without loss of generality, suppose the global optimal is achieved at $(v_1, v_2, \ldots, v_D)$, i.e., $y^* = \mathbb{E}[Y|do(V_1 = v_1, \ldots, V_D = v_D)]$. Note that we have $Pa(Y) \subset \boldsymbol{V}$, which means $\exists V_i \in \boldsymbol{V}$, $V_i \notin Pa(Y)$. For such V_i, the proof is divided into two cases. a). If V_i is independent of Y, denoted as $V_i \perp\!\!\!\perp Y$, then $y^* = \mathbb{E}[Y|do(V_1 = v_1, \ldots, V_{i-1} = v_{i-1}, V_{i+1} = v_{i+1}, \ldots, V_D = v_D)] = \mathbb{E}[Y|do(V_1 = v_1, \ldots, V_D = v_D)]$. b). For another case, if V_i is not independent of Y, given that $V_i \notin Pa(Y)$ and Y is a leaf node in $\mathcal{G}$, thus V_i is an ancestor node of Y. Suppose there are k paths from V_i to Y in $\mathcal{G}$, select a child V^j with $1 \leq j \leq k$ in each path, then Y is conditional independent of V_i conditioning on $\{V^1, V^2, \ldots, V^k\}$, i.e., $V_i \perp\!\!\!\perp Y \,|\, \{V^1, V^2, \ldots, V^k\}$ (this holds because of d-seperation [4]). Since $\{V^1, V^2, \ldots, V^k\} \subset V$, $y^* = \mathbb{E}[Y|do(V_1 = v_1, \ldots, V_D = v_D)] = \mathbb{E}[Y|do(V_1 = v_1, \ldots, V_{i-1} = v_{i-1}, V_{i+1} = v_{i+1}, \ldots, V_D = v_D)]$. In general, from Def. 1, $d \leq D - 1 < D$, the proof is done. □

D Causal Optimization Problems

CGO problem consists of structural causal model $\mathcal{M}$, treatment variable set $\boldsymbol{I}$, intervention range (the optimization domain of treatment variables) $D(\boldsymbol{I})$, target variable Y. We follow the previous work [1] 's methodology to calculate $\mathbb{E}[Y|do(\boldsymbol{X} = \boldsymbol{x})]$ for any specific input $(\boldsymbol{X}, \boldsymbol{x})$. We obtain 1000 samples of Y with all variables in $\boldsymbol{X}$ fixed at $\boldsymbol{x}$, and calculate the mean of these samples to estimate $\mathbb{E}[Y \mid do(\boldsymbol{X} = \boldsymbol{x})]$.

D.1 Real Causal Optimization Problems

Coral Ecology. The work [3] provides an observational dataset but no SCM. Consequently, we construct an SCM using linear regression on the structural equations, denoted by $\boldsymbol{F}$ of SCM, following [1]. As for its $\mathcal{M}$, there are 11 observable variables including seawater nutrition(N), seawater aragonite saturation(O), sea surface chlorophyll(C), seawater alkalinity(T), seawater dissolved inorganic carbon(D), seawater carbon dioxide density(X), Sea Bottom Light Level(L), seawater PH(P), seawater salinity(S), seawater temperature(E). The treatment variables $\boldsymbol{I}$ are $\{L, E, X, S, P\}$. The intervention domain is $D(Nut) = [0, 10]$, $D(\omega_A) = [6, 15]$, $D(Chla) = [-10, 0]$, $D(TA) = [1800, 3000]$, and $D(DIC) = [1750, 2400]$.

Health The problem is from cCBO [2]. The problem is with a defined SCM, which can be written as: Age $= U_{\text{Age}}$; CI $= U_{\text{CI}}$; BMR $= 1500 + 10 \times U_{\text{BMR}}$; Height $= 175 + 10 \times U_{\text{Height}}$; Weight $= \frac{\text{BMR}+6.8\times\text{Age}-5\times\text{Height}}{13.7+\text{CI}\times 150/7716}$; BMI $= \text{Weight}/(\text{Height}/100)^2$; Aspirin $= \sigma(-8.0+0.10\times\text{Age}+0.03\times\text{BMI})$; Statin $= \sigma(-13.0+0.10\times\text{Age}+0.20\times\text{BMI})$; PSA $= 6.8 + 0.04\times\text{Age} - 0.15\times\text{BMI} - 0.04\times\text{Statin} + 0.02\times\text{Aspirin}) + U_{PSA}$, where $U_{\text{Age}} \sim \mathcal{U}(55, 75) U_{\text{CI}} \sim \mathcal{U}(-100, 100), U_{\text{BMR}} \sim t\mathcal{N}(-1, 2), U_{\text{Height}} \sim t\mathcal{N}(-0.5, 0.5), U_{\text{PSA}} \sim t\mathcal{N}(0, 0.4)$. $t\mathcal{N}(a, b)$ means a standard Gaussian random variable truncated between a and b. The intervention domain is $D(\text{CI}) = [0, 100]$, $D(\text{Aspirin}) = [0, 0.2]$, and $D(\text{Statin}) = [0, 0.2]$.

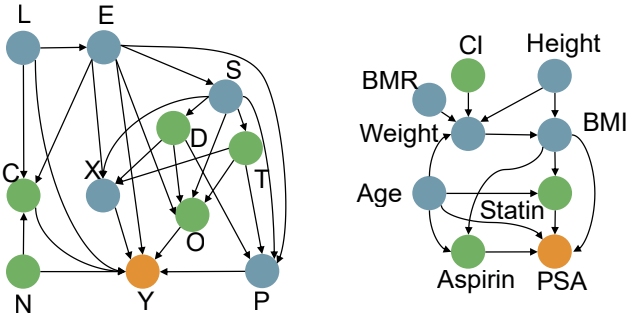

Figure 1. Causal graphs of Coral Ecology and Health. Green nodes refer to treatment variables. Blue nodes mean background variables. Orange nodes are output variables

D.2 Synthetic causal optimization problems

Choices of causal structure model We randomly initialize a structural causal model consisting of $\mathcal{G}$ and $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$. When we input the needed dimension of $\mathcal{M}$, we can get $V \in \boldsymbol{V}$ indexed by 0 to $|\boldsymbol{V}| - 1$. And for each $V_i \in \boldsymbol{V}$ there is U_i influencing on it.

- **Generation of $\mathcal{G}$.** The algorithm to generate a random DAG can be seen in Alg. 3. We adopt probability of a directed connection between two variables p as 0.1.
- **Settings of $P(\boldsymbol{U})$.** We set all $U_i \sim \mathcal{U}(0, 1)$.

- **Generation of $\boldsymbol{F}$.** We consider 3 types of structural equations in synthetic high dimensional problems. (a) Linear: $f_V(Pa(V), U_V) = \sum_{P \in Pa(V)} w_{VP} x_P + b_V + U_V$, w_{VP} and b_V are initialized randomly. There is 50% probability that w_{VP} will be sampled from $\mathcal{U}(0.5, 1)$, while there is another 50% probability that w_{VP} will be sampled from $\mathcal{U}(1, 2)$. $b_V \sim \mathcal{U}(-1, 1)$. $\mathcal{U}(a, b)$ denotes a uniform distribution within the interval (a, b). (b) Additive: $f_V(Pa(V), U_V) = \sum_{P \in Pa(V)} w_{VP} h_{VP}(x_P) + b_V + U_V$. For each h_{VP}, h_{VP} has a 5% probability of being $\sin$, a 5% probability of being $\cos$, a 10% probability of being $\ln$, and an 80% probability of being identity transformation. The initialization of w_{VP} and b_V are same as Linear. (c) Non-Additive: $f_V(Pa(V), U_V) = h_V(Pa(V)) + U_V$. For each V, there is 50% probability that h_V is same as additive structural equations. There is another 50% that $\lfloor |Pa(V)|/6 \rfloor$ non-additive items would be added to the equation. Considering numerical stability, non-additive items are in the form of $\sum_{q=1}^{\lfloor |Pa(V)|/6 \rfloor} x_{V_{lq}} x_{V_{rq}} / |x_{V_{lq}} + x_{V_{rq}}|$, where both V_{lq}, V_{rq} are sampled from $Pa(V)$ uniformly. The final form of the structural equation would be $f_V(Pa(V), U_V) = \sum_{P \in Pa(V)} w_{VP} h_{VP}(x_P) + b_V + \sum_{q=1}^{\lfloor |Pa(V)|/6 \rfloor} x_{V_{lq}} x_{V_{rq}} / |x_{V_{lq}} + x_{V_{rq}}| + U_V$.

Algorithm 3 Construction of a random DAG

Input: Dimension N of Random DAG, probability p of a connection between two variables
Output: Adjacent matrix $\mathcal{G}$ of Random DAG

```
Initialize variable indexes[0, 1, ..., N − 1] in the topological order
for i ∈ [N] do
    for j ∈ {i + 1, i + 2, ..., N − 1} do
        make a directed connection from i to j at the probability of p
    end for
end for
return w after normalized by min-max scaling
```

Choices of treatment variable set We randomly choose $\frac{|\boldsymbol{V}|}{3}$ variables of all variables as treatment variable set $\boldsymbol{I}$.

Choices of intervention bounds We sample 300 observational data D^O to determine the domain of each variable in $\mathcal{M}$. $\max(D^O(I))$ and $\min(D^O(I))$ is utilized to determine the intervention range. $D^O(I)$ means the set of values of variable I in the observational data. We define $l(I) = \max(D^O(I)) - \min(D^O(I))$.there are two ways to determine intervention bounds. One is same strategy where $D(I) = [\min(D^O(I)), \max(D^O(I))]$, another is extension where $D(I) = [\min(D^O(I)) - l(I)/3, \max(D^O(I)) + l(I)]/3$. Both strategies have same probability to be implemented in a specific variable.

E Experiment Details

Baseline experiments are conducted 20 times independently, and the remaining experiments are repeated for 5 times.

E.1 HCBO Details

HCBO uses a Gaussian Process model with additive kernel to deal with CGO problem with linear structural equation types. For other CGO problems, a Gaussian Process model with RBF kernel is used. Average-50 Strategy is utilized to set and update β_1 during the optimization process. β_2 is set as 0.1.

E.2 Implementation Settings

We utilize the author's reference implementations for CBO, MCTS-VS, and TuRBO. The implementation of ALEBO is sourced from the Adaptive Experimentation Platform (Ax). For CMA-ES, we rely on the implementation available in the cmaes package. Additionally, within the BoTorch Library, we have implemented BO, Dropout-BO, REMBO and HCBO. Otherwise, we implement Random Search using Python's random package.

E.3 Parameter Settings

- **CBO** We use the author's default parameter settings.
- **BO** We utilize UpperConfindenceBound acquisition function with $\beta = 0.1$. For CGO problems with linear structural equation types, it use additive kernel to build a GP model. The reason is the same as HCBO. It utilizes RBF in other problems.
- **REMBO** We set the intrinsic embedding dimension $d = \min(|\boldsymbol{I}| - 1, \max(\lfloor |\boldsymbol{I}|/7 \rfloor, \min(20, d_e)))$, d_e is the causal intrinsic dimension found by HCBO method. For CID d_e can be quite small sometimes, we introduce some transformation to improve embedding methods' performance.
- **ALEBO** We set the intrinsic embedding dimension as d set for REMBO.
- **Dropout-BO** We use "copy" as the fill-in strategy for it performs best as the paper's result implies. We set k as the causal intrinsic dimension $\max(2, d)$, d is the intrinsic embedding dimension for REMBO. For inner bo implementation, we use Additive kernel to build a GP model when solving CGO problems with linear structural equation types. We use RBF kernel to solve other problems. It utilize UpperConfindenceBound acquisition function with $\beta = 0.1$.
- **MCTS-VS** We set all parameters as default except for C_p. MCTS-VS paper suggests C_p is set as recommended between 1% and 10% of the optimum of target function. Because we can't know the optimum of a synthetic causal optimization problem before really intervening on it. We use 5% of the best optimization result conducted by HCBO as the C_p.
- **TuRBO** We set the batch size as 20 to suit the more limited cost setting in our problem. Other parameters are set as default settings.
- **CMA-ES** We set the population size as 20, σ as 0.5. Other parameters are set as default.
- **Random Search** We firstly randomly choose a dimension d uniformly. Then we randomly choose a random subset from all subsets with the size d uniformly. Then we choose all values of intervention plan from the intervention domain uniformly.

F CID Validation Experiment

We conduct experiments to confirm whether CID exists in all datasets. Experiments on real-world datasets can be seen in the main paper. We enumerate all subsets of $\boldsymbol{I}$ in real-world datasets and optimize each subset with TuRBO for 1000 iterations. However, in high dimensional settings, enumerating all subsets can be too expensive. We sample some subsets to estimate CID. For synthetic datasets except for Linear-200-66, we sample $\min\{10, C^d_{|\boldsymbol{V}|}\}$ subsets for each dimension d on high dimensional settings to explore the optimum that can be achieved for each dimension. For each subset, we use TuRBO to optimize for 500 iterations. The result can be seen in cf.Fig.2. It indicates that CID always exists in synthetic CGO problems.

Since the number of treatment variables of Linear-200-66 is large, sampling some subsets can still take much computation cost. So we

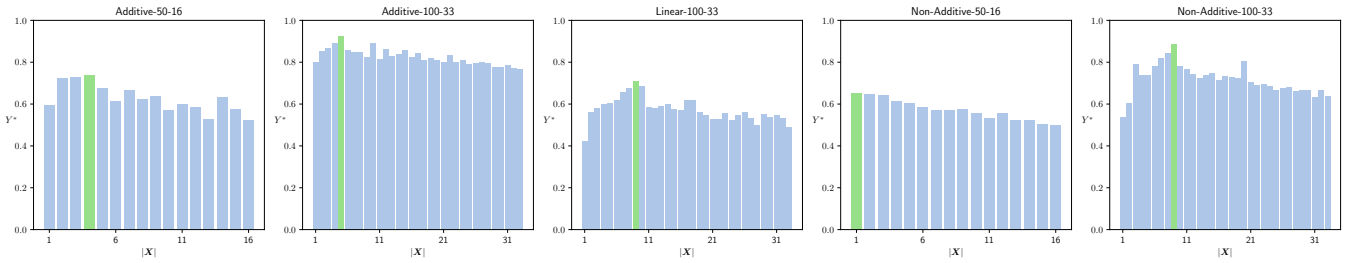

Figure 2. CID validation experiments on synthetic high dimensional problems.

only explore the existence of CID. We optimize the problem for 1000 iterations with TuRBO and obtain optimum as the local optimum that intervened all intervenable variables, $\boldsymbol{I}$. We assume that if its local optimum is worse than the optimum achieved by HCBO, it means that CID exists in this problem. After normalization, optimum of $\boldsymbol{I}$ in Linear-200-66 is 0.6599 which is still smaller than the optimum achieved by HCBO, cf. baseline experiment result in the main paper. The experiment result demonstrates that CID exists in Linear-200-66.

G Hyperparameter Analysis

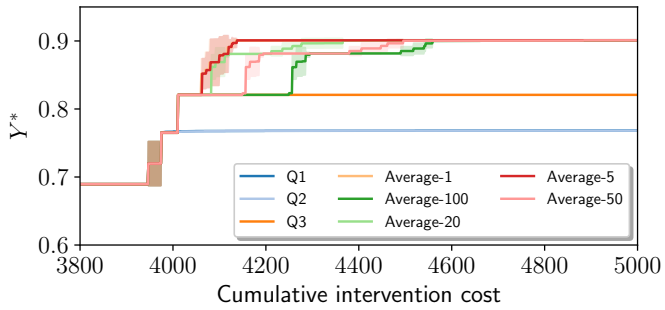

Figure 3. β_1 hyperparameter analysis about static and dynamic strategies.

$$ISSF(\boldsymbol{X}) = \bar{Y}_{D^I_{\boldsymbol{X}}} + \beta_1 UCB(\boldsymbol{x}^*_{\boldsymbol{X}}|GP_{\tilde{D}^I_{\boldsymbol{X}}}) \,. \tag{1}$$

The hyperparameter experiment is conducted in Linear-100-33 dataset to investigate optimal selection strategy of β_1, cf. Eq. (1), in ISSF component. We identify a reasonable search domain of β_1, and compare two different strategies: dynamic and static.

The search domain D_{β_1} can be seen in Eq. (2). Since β_1 is designed to balance exploration $UCB(\boldsymbol{x}^*_{\boldsymbol{X}}|GP_{\tilde{D}^I_{\boldsymbol{X}}})$ and exploitation $\bar{Y}_{D^I_{\boldsymbol{X}}}$ at the same level, the search domain D_{β_1} of β_1 can be derived from the quality transformation of Eq. (1).

$$D_{\beta_1} = \{\frac{\bar{Y}_{max} - \bar{Y}_{min}}{UCB(\boldsymbol{x}^*_{\boldsymbol{X}}|GP_{\tilde{D}^I_{\boldsymbol{X}}})} | \boldsymbol{X} \in \text{ECCIS}(\mathcal{M})\} \tag{2}$$

Two strategies are defined as below. Static strategy fixes β_1 at the first iteration while dynamic strategy updates β_1 for every k iterations. For static strategy, we consider the three quartiles of D_{β_1}, denoted as Q1, Q2, Q3. For dynamic strategy, we compare Average-k and Median-k. The former one update β_1 with average value of D_{β_1}, while the latter one uses the median value. k can be among $1, 5, 20, 50, 100$.

The comparison result can be seen in Fig.3. Because the result of Median-k is close to that of Average-k, we only show the static strategy and dynamic Average strategy. From Fig.3, we can observe that frequent updates on β_1 can help reach better optimum. Dynamic method's superiority can be attributed to the fact that magnitude difference between $\bar{Y}_{D^I_{\boldsymbol{X}}}$ and $UCB(\boldsymbol{x}^*_{\boldsymbol{X}}|GP_{\tilde{D}^I_{\boldsymbol{X}}})$ changes during the optimization process. And the Fixed methods can't handle it. One thing to note here is that all dynamic strategies locate optimum successfully, which shows that the parameter k is robust in this situation. Although the result suggests that more frequent updates of β_1 lead to quicker convergence in optimization, we adopt a more conservative strategy Average-50. Our consideration is that theoretically, excessively frequent updates on β_1 may make ISSF focus too much on exploration rather than exploitation. As the result shown in other experiments, Average-50 still achieves SOTA performance.

H Additional Analysis of Baseline Experiment

We analyze the effectiveness of different baseline methods in CGO problems, based on our observations, cf. Fig. 2 in Sec. 5 of the main paper.

- Random Search often ranks as the second-best method for finding optimal solutions across various problems. It always starts at a low level then gradually surpasses standard optimization methods. This reflects that no matter how strong a standard optimization method is, it can't locate global optimum without utilizing the causal intrinsic dimension information. Intervening on all variables may not achieve optimum.
- TuRBO, Dropout-BO and MCTS-VS perform well among standard optimization method. TuRBO can adjust the trust region to locate optimum. Dropout-BO and MCTS-VS optimize on selected variables rather than all variables. Both features accelerate their convergence. But they can't detect the optimum for they can only intervene on entirety of $\boldsymbol{I}$.
- Embedding methods(REMBO and ALEBO) don't perform well in non-liner settings. This can be attributed to this setting's violation to intrinsic linear embedding assumptions. Notably, in Coral Ecology dataset, ALEBO converges faster. This is because Coral Ecology follows linear structural equation type, which aligns with the linear embedding assumptions required by ALEBO.
- Limited by budget, CMA-ES fails in all problems. Moreover, CMA-ES needs some tunings to ensure performance in CGO scenarios.

We also conduct statistical tests on the final optimum results of baseline experiments, cf. Tab. 2

Table 2. Comparison of HCBO with baseline methods across different problems. Each cell indicates whether HCBO outperforms another baseline method according to t-tests on the final optimum for the corresponding problem. For each t-test comparison, we assumed that the HCBO method performed better and set the p-value threshold to 0.1. A checkmark (✓) signifies that HCBO outperforms the baseline, while a cross (×) indicates it does not.

	CBO	Random-Search	CMA-ES	REMBO	ALEBO	MCTS-VS	Dropout-BO	TuRBO	BO
Health	✓	✓	✓	✓	✓	✓	✓	✓	✓
Coral Ecology	✓	✓	✓	✓	×	✓	×	×	✓
Linear-100-33	✓	✓	✓	✓	✓	✓	✓	✓	✓
Linear-200-66	✓	✓	✓	✓	✓	✓	✓	✓	✓
Additive-50-16	✓	✓	✓	✓	✓	✓	✓	✓	✓
Additive-100-33	✓	✓	✓	✓	✓	✓	✓	✓	✓
Non-Additive-50-16	✓	✓	✓	✓	✓	✓	✓	✓	✓
Non-Additive-100-33	✓	✓	✓	✓	✓	✓	✓	✓	✓

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)*, volume 108, pages 3155–3164, Sicily, Italy, 2020.

[2] V. Aglietti, A. Malek, I. Ktena, and S. Chiappa. Constrained causal bayesian optimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 304–321, Honolulu, Hawaii, USA, 2023.

[3] 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.

[4] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search (Second Edition)*. MIT Press, 2000.