MODEL-BASED CAUSAL BAYESIAN OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

How should we intervene on an unknown structural equation model to maximize a downstream variable of interest? This setting, also known as causal Bayesian optimization (CBO), has important applications in medicine, ecology, and manufacturing. Standard Bayesian optimization algorithms fail to effectively leverage the underlying causal structure. Existing CBO approaches assume noiseless measurements and do not come with guarantees. We propose the *model-based causal Bayesian optimization algorithm (*MCBO*)* that learns a full system model instead of only modeling intervention-reward pairs. MCBO propagates epistemic uncertainty about the causal mechanisms through the graph and trades off exploration and exploitation via the optimism principle. We bound its cumulative regret, and obtain the first non-asymptotic bounds for CBO. Unlike in standard Bayesian optimization, our acquisition function cannot be evaluated in closed form, so we show how the reparameterization trick can be used to apply gradient-based optimizers. The resulting practical implementation of MCBO compares favorably with state-of-the-art approaches empirically.

1 INTRODUCTION

Many applications, such as drug and material discovery, robotics, agriculture, and automated machine learning, require optimizing an unknown function that is expensive to evaluate. *Bayesian optimization (BO)* is an efficient framework for sequential optimization of such objectives (Močkus, 1975). The key idea in BO is to quantify uncertainty in predictions via a probabilistic model and use the uncertainty to negotiate the exploration–exploitation dilemma. While most standard BO methods focus on a black-box setup (Figure 1 a), in practice, we often have more structure that we can exploit to improve sample efficiency.

In this paper, we *exploit structural knowledge in the form of a causal model*. In particular, we assume that actions can be modeled as interventions on a structural causal model (SCM) (Pearl, 2009) that contains the reward (function output) as a variable (Figure 1 b). While we assume the graph structure to be known, we consider the *functional relations* in the SCM as unknown. All variables in the SCM are observed along with the reward after each round of interaction. This Causal BO setting has important potential applications, such as optimizing medical and ecological interventions (Aglietti et al., 2020b). For illustrative purposes, consider an example of an agronomist trying to find the optimal Nitrogen fertilizer schedule for maximizing crop yield, described in Figure 1. There, the concentration of Nitrogen in the soil causes its concentration in the soil at the later timesteps.

To exploit this causal graph structure for optimization, we propose the *model-based causal Bayesian optimization algorithm* (MCBO). MCBO explicitly models the full SCM and the accompanying uncertainty of all SCM components. This allows our algorithm to select interventions based on an optimistic strategy similar to that used by the upper confidence bound algorithm (Srinivas et al., 2010). We show that this strategy leads to the first CBO algorithm with a cumulative regret guarantee. For a practical algorithm, maximizing the upper confidence bound in our setting is computationally more difficult, because system uncertainty must be propagated through the entire estimated SCM to the reward variable. We show that an application of the reparameterization trick allows MCBO to be practically implemented with common gradient-based optimizers. Empirically, we demonstrate that MCBO achieves competitive performance on existing CBO benchmarks and a related setting called function network BO (Astudillo & Frazier, 2021).

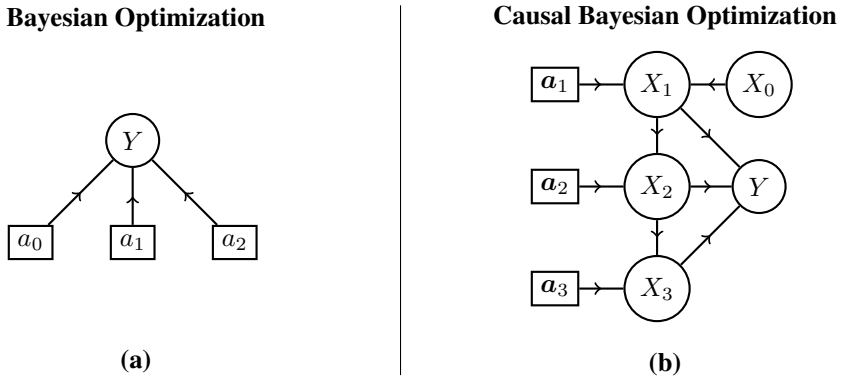

Figure 1: **(a)** the DAG used in Bayesian optimization, regardless of the problem structure. **(b)** An example of a DAG in causal Bayesian optimisation. Circular nodes are observations and square nodes are action inputs that correspond to an unknown soft intervention on the child node. Both Bayesian optimization and causal Bayesian optimization select $\boldsymbol{a}$ and then observe variables in circular nodes ($\boldsymbol{X}$ and Y). Each observation is modelled with its own Gaussian process. This DAG is used in our stylised agronomy example, where the goal is to maximize crop yield Y. X_0 is an unmodifiable starting property of the soil, and $X_1, \dots X_3$ are the measured amounts of Nitrogen in the soil at different timesteps. $\mathbf{a}_1 \dots \mathbf{a}_3$ are possible interventions of adding synthetic fertilizer to the soil.

Contributions

- We introduce MCBO, a model-based algorithm for causal Bayesian optimization than can be applied with very generic classes of interventions.
- We prove the first sublinear cumulative regret bound of MCBO, and show how it scales in terms of the causal graph structure and a notion of the maximum information gain of the model class.
- By an application of the reparameterization technique, we show how our algorithm can be efficiently implemented with popular gradient-based optimizers.
- We evaluate MCBO on existing hard-intervention CBO tasks and a related setting on function networks. Our results show that MCBO performs favorably compared to methods designed specifically for these tasks.

2 Background and Problem Statement

Structure We consider a structural equation model consisting of a known directed acyclic graph $\mathcal{G}$ (DAG) and a tuple $\langle Y, \mathbf{X}, \boldsymbol{f} \rangle$, where Y is the reward variable to be maximized, $\mathbf{X} = \{X_i\}_{i=0}^{m-1}$ is a set of observed random variables, and the set $\boldsymbol{f} = \{f_i\}_{i=0}^{m}$ defines the functional relations between these variables. We use Y and X_m interchangeably. We assume the Markov property for our causal model (Pearl, 2009) and no unobserved confounding. Therefore, the joint observational distribution can be factored as $\mathbb{P}(\mathbf{X}) = \prod_i \mathbb{P}(X_i \mid Z_i)$, where $Z_i = \{X_j\}_{j \in pa_i}$ denotes the set of variables in $\mathbf{X}$ that are parent variables of X_i in $\mathcal{G}$, i.e., $pa_i \subset [0, \dots, m]$. The $\mathbf{X}$ are ordered according to a valid topological ordering (so X_0 is a root and X_m is a sink).

Each observation $X_i \in \mathcal{X}_i \subset \mathbb{R}^{d_i}$ is defined over a compact set for all $i \in [0, \dots, m-1]$. We define the space of all non-reward variables with $\mathcal{X} = \cup_{i=0}^{m-1} \mathcal{X}_i$. The reward is a scalar $X_m \in \mathbb{R}$. We denote the set of all possible values for the parents of X_i as $\mathcal{Z}_i = \cup_{j \in pa_i} \mathcal{X}_j$. Each X_i is generated according to the function $f_i : \mathcal{Z}_i \to \mathcal{X}_i$, taking the parent nodes Z_i of X_i as input: $X_i = f_i(Z_i) + \boldsymbol{\omega}_i$ where $\boldsymbol{\omega}_i$ is zero-mean noise with known distribution. While $\mathcal{G}$ is known, the elements of $\boldsymbol{f}$ are unknown but can be learned by inputting actions into the system.

Actions as interventions We model the actions as *interventions* on the system described above and parameterize them by controllable variables $\boldsymbol{a}$ as follows. Let $\mathcal{A}_i \subset \mathbb{R}^q$ denote the compact space of

each action $\boldsymbol{a_i}$ and $\mathcal{A}$ be the space of all actions $\boldsymbol{a} = \{\boldsymbol{a}_i\}_{i=0}^m$. Specifically, we represent actions as additional nodes in $\mathcal{G}$: $\boldsymbol{a}_i$ is a parent of X_i if and only if $i = j$ and an additional input to f_i. This is known as a *soft intervention with unknown effect* and it describes a very general class of interventions (Pearl, 2009). A simple example of a soft intervention is a shift intervention, which results in $\boldsymbol{x}_i = f_i(\boldsymbol{z}_i, \boldsymbol{a}_i) + \boldsymbol{\omega}_i = g_i(\boldsymbol{z}_i) + \boldsymbol{a}_i + \boldsymbol{\omega}_i$ for some function g_i. A shift intervention might occur in our example of adding Nitrogen fertilizer to soil and then measuring the total soil Nitrogen concentration.

In contrast to soft interventions, so-called hard interventions (often referred to as do-interventions) modify the targeted variable to a specific distribution, independent of the variable's parents. Hard interventions therefore also modify the graph structure. Finally, one can also obtain observational data when no intervention is performed on the system. In our notation, not intervening on node i is equivalent to setting $\boldsymbol{a}_i = 0$. An example would be not applying any Nitrogen fertilizer to the soil. In this paper, while we mainly focus on soft interventions, we also consider observational data and hard interventions in the experiments.

Constraints In many applications, there exist several nodes X_i that cannot be intervened upon at all. For example, the soil starting conditions in Figure 1 (b). In that case, we fix the corresponding $\boldsymbol{a_i}$ to $\mathbf{0}$. Often, among intervenable nodes, we cannot intervene on all simultaneously. For example, a farmer may only have capacity to apply fertilizer at 2 out of 3 of the possible time windows. The example of a cardinality constraint can be written as:

$$\mathcal{A} = \left\{ \boldsymbol{a} = \{\boldsymbol{a}_i\}_{i=0}^m : \sum_{i=0}^m \|\boldsymbol{a}_i\|_0 \leq c, c \geq 1 \right\}. \tag{1}$$

Our theoretical results hold for arbitrary compact constraint sets $\mathcal{A}$.

Problem statement We consider the problem of sequentially interacting with the system described by a fixed but unknown set of functions $\boldsymbol{f}$ with $f_i : \mathcal{Z}_i \times \mathcal{A}_i \to \mathcal{X}_i$. The learner selects an intervention $\boldsymbol{a} = \{\boldsymbol{a}_i\}_{i=0}^m$, and obtains a set of noisy observations for all the nodes in $\mathcal{G}$:

$$\boldsymbol{x}_i = f_i(\boldsymbol{z}_i, \boldsymbol{a}_i) + \boldsymbol{\omega}_i, \;\; \forall i \in [0, \dots, m], \tag{2}$$

If $\boldsymbol{x}_i$ is a root node, Z_i will be an empty vector and f_i only depends on $\boldsymbol{a}_i$. Since we cannot intervene on the target variable X_m, we fix $\boldsymbol{a}_m = 0$. This results in the reward evaluation:

$$y = f_m(\boldsymbol{z}_m, \boldsymbol{a}_m) + \boldsymbol{\omega}_m, \tag{3}$$

that implicitly depends on the whole intervention $\boldsymbol{a}$ through the realizations of its parents. For simplicity, we omit the time dependence in the observation. In particular, the learner sequentially observes evaluations $\boldsymbol{x}_t = \{\boldsymbol{x}_{i,t}\}_{i=0}^m$ for the chosen actions $\boldsymbol{a}_t = \{\boldsymbol{a}_{i,t}\}_{i=0}^m$.[1] Our objective is to select $\boldsymbol{a}$ to maximize the expected reward:

$$\boldsymbol{a}^* = \arg\max_{\boldsymbol{a} \in \mathcal{A}} \mathbb{E}[y|\boldsymbol{a}], \tag{4}$$

where, unless otherwise stated, expectations are taken over noise $\boldsymbol{\omega} = [\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_m]$.

Performance metric Our goal is to design a sequence of interventions $\boldsymbol{a}_t$ that eventually maximizes the expected reward Eq. (4). We study this convergence via the *cumulative regret* over a time horizon T :

$$R_T = \sum_{t=1}^T \Big[\mathbb{E}[y|\boldsymbol{a}^*] - \mathbb{E}[y|\boldsymbol{a}_t] \Big]. \tag{5}$$

A sublinear growth rate of R_T with T implies vanishing average regret $R_T/T \to 0$ as $T \to \infty$. Intuitively, this implies the existence of some t after which we are arbitrarily close to the optimal value. As an alternative to cumulative regret, one can also study convergence through other notions such as simple regret. The most appropriate metric depends on the application. In the Nitrogen fertilizer example, cumulative regret might be preferable because we care about getting high crop yields across all years of Nitrogen tuning, not just in only one year.

[1] Here, by i we index the nodes $[0, \dots, m]$ in graph and by t we enumerate the rounds of interacting with the system.

Regularity assumptions. We consider standard smoothness assumptions for the unknown functions $f_i : \mathcal{S} \to \mathcal{X}_i$ defined over some compact domain $\mathcal{S}$ (Srinivas et al., 2010). In particular, for each node $i \in [0, \dots, m]$, we assume that $f_i(\cdot)$ belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_{k_i}$, a space of smooth and vector-valued functions defined on $\mathcal{S} = \mathcal{Z}_i \times \mathcal{A}_i$. This means that $f_i \in \mathcal{H}_{k_i}$ is induced by a vector-valued kernel function $k_i : \mathcal{S} \times \mathcal{S} \to \mathcal{X}_i$. We also assume that $k_i(s, s') \leq \mathbf{1}$ for every $s, s' \in \mathcal{S}$. Moreover, the RKHS norm of $f_i(\cdot)$ is assumed to be bounded $\|f_i\|_{k_i} \leq \mathcal{B}_i$ for some fixed constant $\mathcal{B}_i > 0$. Finally, we also assume that the noise is bounded, i.e., $\boldsymbol{\omega}_i \in [-1, 1]^{d_i}$. This ensures the compactness of the domains $\mathcal{Z}_i$.[2]

Hard interventions MCBO can be also applied in a setup with hard interventions. In this case, it can be convenient to adapt the notation. Instead of selecting $\boldsymbol{a}$, the algorithm must select a set of intervention targets $I \in \mathcal{I} \subset \mathcal{P}(\{0, 1, ..., m-1\})$ from a set of all possible subsets $\mathcal{P}$ and their values $\boldsymbol{a}_I \in \mathcal{A}_I$. Eq. (2) then takes the form of

$$\boldsymbol{x}_i = \begin{cases} \boldsymbol{a}_i & \text{if } i \in I \\ f_i(\boldsymbol{z}_i) + \boldsymbol{\omega}_i & \text{otherwise,} \end{cases} \qquad \forall\, i \in [0, \dots, m]. \tag{6}$$

Here f_i is still unknown and modeled using the same regularity assumptions.

2.1 Related Work

Standard BO approaches model only Y given $\boldsymbol{a}$, as shown in Figure 1. Aglietti et al. (2020b) introduce the original CBO setting, but only for the setting of hard interventions. Aglietti et al. (2020a) propose a method that generalizes across data on all intervention targets when estimating Y given $\boldsymbol{a}$, but does not learn an explicit model over all variables. Both works consider unobserved confounding and apply the do-calculus (Pearl, 2009) to estimate Y given $\boldsymbol{a}$ in this case. Branchini et al. (2022) and Alabed & Yoneki (2022) extend these works to explore the CBO framework in the case of unknown $\mathcal{G}$. Lattimore et al. (2016) introduce a setup for optimization over interventions in an SCM, but for the bandit setting with binary random variables.

Function network BO (Astudillo & Frazier, 2021) is similar to the setting of MCBO but involves no additive noise: the system is entirely deterministic. Moreover, function networks are not causal models and do not model actions as interventions. Kusakawa et al. (2021) study function networks in the special case of a chain graph.

The function class in this paper is similar to deep Gaussian processes (GPs) (Damianou & Lawrence, 2013) in that MCBO models Y as a composition of GPs given $\boldsymbol{a}$. Deep GPs, however, do not make use of intermediate system variables and do not compose GPs according to a causal graph structure.

Our use of the reparameterization trick to practically implement an upper confidence bound acquisition function (Srinivas et al., 2010) in MCBO is inspired by Curi et al. (2020), who apply ideas from BO for sample efficient reinforcement learning.

3 Algorithm

Statistical model We use GPs for learning the RKHS functions $f_0, \cdots f_m$ from observations. Our regularity assumptions permit the construction of confidence bounds using these GP models with priors associated with the RKHS kernels. We refer to Rasmussen (2003) for more background on the relation between GPs and RKHS functions. Let $\boldsymbol{\mu}_{i,0}$ and $\boldsymbol{\sigma}^2_{i,0}$ denote the prior mean and variance functions for each f_i, respectively. Then, the corresponding posterior GP mean and variance denoted by $\boldsymbol{\mu}_{i,t}(\cdot)$ and $\boldsymbol{\Sigma}^2_{i,t}(\cdot)$, respectively, are computed based on the previous evaluations $\mathcal{D}_t = \{\boldsymbol{z}_{1:t}, \boldsymbol{a}_{1:t}, \boldsymbol{x}_{1:t}\}$, where each element includes the evaluation over the whole graph $\mathcal{G}$, e.g., $\boldsymbol{x}_{1:t} = \{\boldsymbol{x}^i_{1:t}\}_{i=0}^m$, $\boldsymbol{x}^i_{1:t} = [\boldsymbol{x}_{i,1}, \dots, \boldsymbol{x}_{i,t}]$.[3] In particular, for each function $f_i(\cdot)$

[2]This assumption can be relaxed to $\boldsymbol{\omega}_i$ being σ_i-sub-Gaussian using similar techniques to Curi et al. (2020) Appendix I. Though sub-Gaussian noise includes distributions with unbounded support, Curi et al. (2020) provide high probability bounds on the domain of $\mathcal{Z}_i$.

[3]We refer to Appendix A.1.1 for more details on the notation.

belonging to RKHS defined by the given kernel $k_i(\cdot,\cdot)$:

$$\boldsymbol{\mu}_{i,t}(\boldsymbol{z}_i, \boldsymbol{a}_i) = \boldsymbol{k}_{i,t}(\boldsymbol{z}_i, \boldsymbol{a}_i)^\top (\boldsymbol{K}_t + \sigma_i^2 \boldsymbol{I})^{-1} \boldsymbol{x}_{1:t}^i, \tag{7}$$

$$\boldsymbol{\sigma}_{i,t}^2(\boldsymbol{z}_i, \boldsymbol{a}_i) = k_i(\boldsymbol{z}_i, \boldsymbol{a}_i; \boldsymbol{z}_i, \boldsymbol{a}_i) - \boldsymbol{k}_t(\boldsymbol{z}_i, \boldsymbol{a}_i)^\top (\boldsymbol{K}_t + \sigma_i^2 \boldsymbol{I})^{-1} \boldsymbol{k}_t(\boldsymbol{z}_i, \boldsymbol{a}_i), \tag{8}$$

where $(\boldsymbol{K}_t)_{i,j} = k_i(\boldsymbol{z}_i, \boldsymbol{a}_i; \boldsymbol{z}_j, \boldsymbol{a}_j)$, $\boldsymbol{k}_{i,t}(\boldsymbol{z}, \boldsymbol{a})^\top = [k_i(\boldsymbol{z}_1, \boldsymbol{a}_1; \boldsymbol{z}, \boldsymbol{a}), \ldots, k_i(\boldsymbol{z}_t, \boldsymbol{a}_t; \boldsymbol{z}, \boldsymbol{a})]^\top$, the identity matrix is denoted by $\boldsymbol{I}$, and σ_i is the variance proxy for $\boldsymbol{\omega}_i$. At time t, the known set $\mathcal{M}_t$ of statistically plausible functions $\tilde{\boldsymbol{f}} = \{\tilde{f}_i\}_{i=0}^m$ that lie inside the confidence interval is defined by the posterior of the GP:

$$\mathcal{M}_t = \left\{ \tilde{\boldsymbol{f}} \;\middle|\; \forall i : \left| \tilde{f}_i(\boldsymbol{z}_i, \boldsymbol{a}_i) - \boldsymbol{\mu}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i) \right| \leq \beta_{i,t} \boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i),\ \forall \boldsymbol{z}_i \in \mathcal{Z}_i, \boldsymbol{a}_i \in \mathcal{A}_i \right\}. \tag{9}$$

Here, $\beta_{i,t}$ is a parameter that ensures the validity of the confidence bounds. Some examples of such concentration inequalities under similar regularity assumptions as well as explicit forms for $\beta_{i,t}$ can be found in Chowdhury & Gopalan (2017) and Srinivas et al. (2010). In the following, we set $\beta_{i,t} = \beta_t$ for all i such that the confidence bounds are still valid.

Acquisition function At each round t, interventions are selected based on maximizing an acquisition function. We use an upper confidence bound acquisition function, which is defined as:

$$\boldsymbol{a}_t = \arg\max_{\boldsymbol{a} \in \mathcal{A}} \max_{\tilde{\boldsymbol{f}} \in \mathcal{M}_t} \mathbb{E}_{\boldsymbol{\omega}} \Big[y \mid \tilde{\boldsymbol{f}}, \boldsymbol{a} \Big]. \tag{10}$$

Thus, we optimistically pick interventions that yield the highest expected return among all functions that are still plausible given past observations. Eq. (10) is not amenable to commonly used optimization techniques, due to the maximization over a set of functions. Therefore, following Curi et al. (2020), we use the reparameterization trick to write any $\tilde{f}_i \in \tilde{\boldsymbol{f}} \in \mathcal{M}_t$ using some function $\eta_i : \mathcal{Z}_i \times \mathcal{A}_i \to [-1, 1]^{d_i}$ as:

$$\tilde{f}_{i,t}(\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{a}}_i) = \boldsymbol{\mu}_{i,t-1}(\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{a}}_i) + \beta_{i,t} \boldsymbol{\Sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{a}}_i) \eta_i(\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{a}}_i), \tag{11}$$

where observations passed through the plausible model, and not necessarily the true model $\mathbf{f}$, are denoted as $\tilde{\boldsymbol{x}}_i = \tilde{f}_i(\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{a}}_i) + \tilde{\boldsymbol{\omega}}_i$. This allows for rewriting our acquisition function as optimisation over $\boldsymbol{\eta} : \mathcal{Z} \times \mathcal{A} \to [-1, 1]^{|\mathcal{X}|}$:

$$\boldsymbol{a}_t = \arg\max_{\boldsymbol{a} \in \mathcal{A}} \max_{\boldsymbol{\eta}(\cdot)} \mathbb{E}_{\boldsymbol{\omega}} \Big[y \mid \tilde{\boldsymbol{f}}, \boldsymbol{a} \Big], \quad \text{s.t. } \tilde{\boldsymbol{f}} = \{\tilde{f}_i\}_i \text{ in Eq. (11).} \tag{12}$$

Intuitively, the variables $\boldsymbol{\eta}$ allow for choosing optimistic but plausible models given the confidence bounds. The $\boldsymbol{\eta}$ can be written as a parametric function, and then standard optimization techniques applied. For the theory, we assume access to an oracle that can find the global optimum for Eq. (12). In practice, such an oracle may be computationally infeasible due to the non-convexity of Eq. (12). We discuss the heuristics that we use for approximating this oracle in the appendix.

Algorithm 1 summarizes our Model-based Causal BO approach. We note that, for the special case of the SCM following the DAG of Fig. 1, our algorithm and the associated guarantees reduce to standard BO (cf. Srinivas et al., 2010).

Hard interventions MCBO also naturally generalizes to hard interventions, cf. Algorithm 2. In our experiments, we perform the combinatorial optimisation over $\mathcal{I}$ by enumeration, because $|\mathcal{I}|$ is not large for the instances we consider. We apply the method of Lee & Bareinboim (2019) to prune sets of intervention targets that contain interventions, resulting in a smaller set to optimize over.

Algorithm 1 UCB Model-based Causal BO (MCBO)

Require: Parameters $\{\beta_t\}_{t\geq 1}, \{\sigma_i\}_i$, Prior means $\{\boldsymbol{\mu}_{i,0} = 0\}_i$, Kernel functions $\{k_{i,0}\}_i$
1: **for** $t = 1, 2, \ldots$ **do**
2: Construct confidence bounds as in Eq. (9)
3: Select $\boldsymbol{a}_t \in \arg\max_{\boldsymbol{a}\in\mathcal{A}} \max_{\boldsymbol{\eta}(\cdot)} \mathbb{E}[y \mid \{\tilde{f}_i\}, \boldsymbol{a}]$ as in Eq. (12)
4: Observe samples $\{\boldsymbol{z}_{i,t}\}_i$, $\{\boldsymbol{x}_{i,t}\}_i$ and y_t
5: Use $\mathcal{D}_t$ to update posterior $\{\boldsymbol{\mu}_{i,t}(\cdot)\}$ and $\{\boldsymbol{\sigma}^2_{i,t}(\cdot)\}_i$ as in Eqs. (7) and (8)
6: **end for**

Algorithm 2 UCB Model-based Causal BO with Hard Interventions (MCBO)

Require: Parameters $\{\beta_t\}_{t\geq 1}, \{\sigma_i\}_i$, Prior means $\{\boldsymbol{\mu}_{i,0} = 0\}_i$, Kernel functions $\{k_{i,0}\}_i$
1: **for** $t = 1, 2, \ldots$ **do**
2: Construct confidence bounds as in Eq. (9)
3: Select $I, \boldsymbol{a}_I \in \arg\max_{I,\boldsymbol{a}_I} \max_{\boldsymbol{\eta}} \mathbb{E}[y \mid \{\tilde{f}_i\}, do(X_I = \boldsymbol{a}_I)]$
4: Observe samples $\{\boldsymbol{z}_{i,t}\}_i$, $\{\boldsymbol{x}_{i,t}\}_i$ and y_t
5: Use $\mathcal{D}_t$ to update posterior $\{\boldsymbol{\mu}_{i,t}(\cdot)\}$ and $\{\boldsymbol{\sigma}^2_{i,t}(\cdot)\}_i$ as in Eqs. (7) and (8)
6: **end for**

4 Theoretical Analysis

This section describes the convergence guarantees for MCBO, showing the first sublinear cumulative regret bounds for causal BO. We start by introducing the assumptions required for this analysis.

Assumption 1 (*Functional relations*)**.** All $f_i \in f$ are L_f-Lipschitz continuous and, for all $i \geq 0$, the elements of the noise vector ω_i are *i.i.d.* σ_i-sub-Gaussian.

Assumption 2 (*Continuity*)**.** For all i, the functions $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are L_μ and L_σ Lipschitz continuous.

Assumption 3 (*Calibrated model*)**.** All statistical models are *calibrated* w.r.t. f, so that with $\boldsymbol{\sigma}_{i,t}(\cdot)$ there exists a sequence $\beta_t \in \mathbb{R}_{>0}$ such that, with probability at least $(1-\delta)$, it holds jointly for all i, and $t \geq 0$ and $\boldsymbol{z}_i, \boldsymbol{a}_i \in \mathcal{Z}_i \times \mathcal{A}_i$ that $|f_i(\boldsymbol{z}_i, \boldsymbol{a}_i) - \boldsymbol{\mu}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)| \leq \beta_t \boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)$, element-wise.

Assumption 1 follows directly from assuming bounded noise and the regularity assumptions. Assumption 2 holds if the RKHS of each f_i has a Lipschitz continuous kernel (see Curi et al. (2020), Appendix G). This means that Assumption 2 restricts the convergence guarantees to soft interventions that affect their target variable in a smooth way, meaning that our analysis does not apply to the case of hard interventions. We nevertheless experimentally demonstrate the effectiveness of MCBO in non-smooth settings, such as CBO with hard interventions. Assumption 3 holds when we assume that the ith GP prior uses the same kernel as the RKHS of f_i and that β_t is sufficiently large to ensure the confidence bounds in Eq. (9) hold.

In the DAG $\mathcal{G}$ over nodes $\{X_i\}_{i=0}^m$, let N denote the maximum distance from a root node to X_m: $N = \max_i \mathrm{dist}(X_i, X_m)$ where $\mathrm{dist}(\cdot,\cdot)$ is measured as the number of edges in the longest path from a node X_i to the reward node X_m. Let K denote the maximum number of parents of any variable in $\mathcal{G}$. The following theorem bounds the performance of MCBO in terms of cumulative regret.

Theorem 1. *Consider the optimization problem in Eq.* (4) *with known causal graph $\mathcal{G}$ and unknown functional relations Eq.* (2) *that satisfy Assumptions 1–3. Let $\{\boldsymbol{a}_t \in \mathcal{A}, \boldsymbol{z}_t \in \mathcal{Z}\}_{t>0}$ denote the set of actions chosen by* MCBO *and the parent variable realizations, respectively. Then, provided $\mathcal{A}_i$ is compact $\forall i$, for all $T \geq 1$, with probability at least $1-\delta$, the cumulative regret of Algorithm 1 is bounded by*

$$R_T \leq \mathcal{O}\Big(L_f^N L_\sigma^N \beta_T^N K^N m \sqrt{T\,\gamma_T}\Big)\,.$$

Here, $\gamma_T = \max_i \gamma_{i,T}$ where the node-specific $\gamma_{i,T}$ denotes the *maximum information gain* at a time T commonly used in the standard regret guarantees for Bayesian optimization (Srinivas et al., 2010). This maximum information gain is known to be sublinear in T for many common kernels,

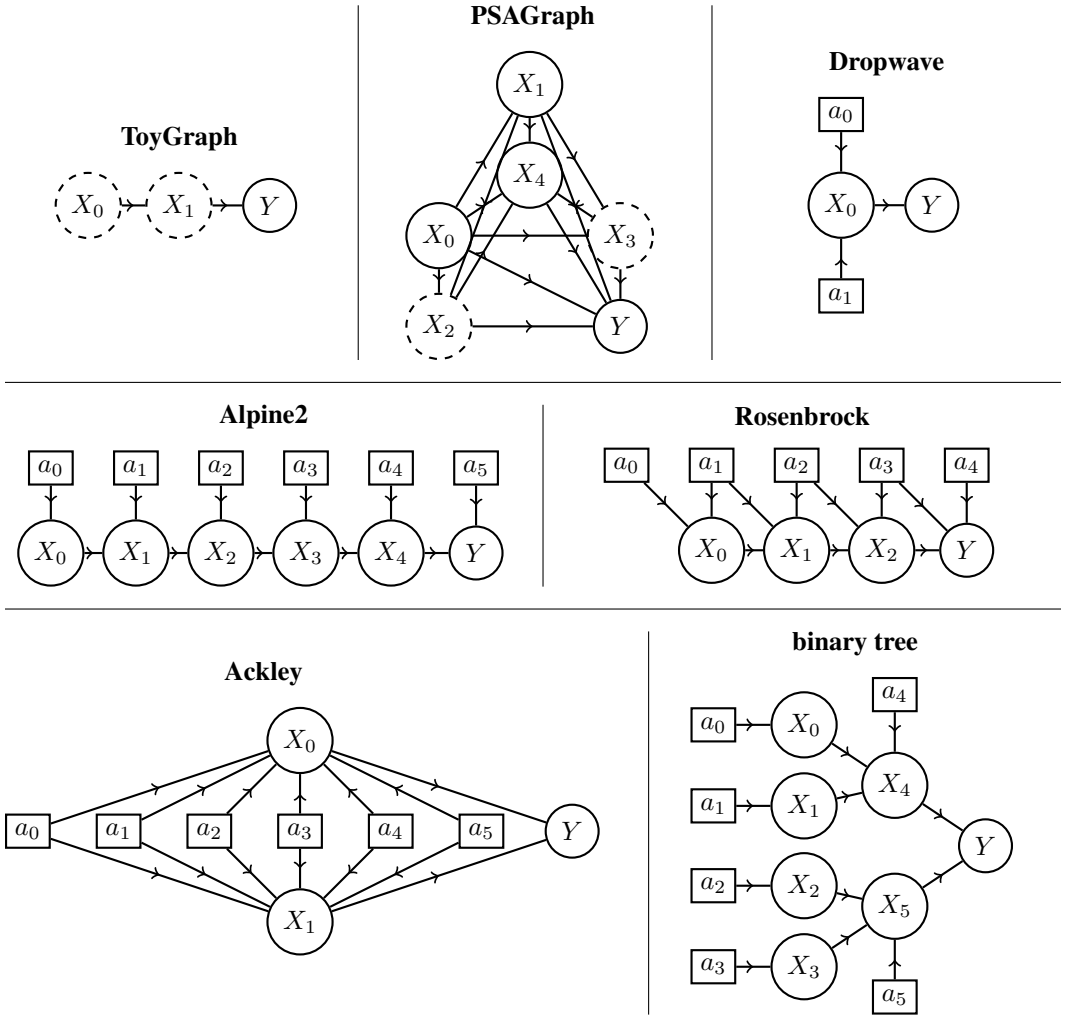

Figure 2: The DAGs corresponding to each task in the experiments, except binary tree which is used for illustrative purposes. Circles are observation or reward variables. Squares are actions in the function network setting. In hard intervention CBO, nodes with a dashed border can be intervened upon. All nodes represent a scalar random variable.

such as linear and squared exponential kernels, resulting in an overall sublinear regret for MCBO. We refer to Appendix A.1.4 for the proof.

Theorem 1 demonstrates that the use of the graph structure in MCBO results in a potentially *exponential* improvement in how the cumulative regret scales with the number of actions m. Standard Bayesian optimization as in Figure 1 (a), that makes no use of the graph structure, results in cumulative regret exponential in m (Srinivas et al., 2010), when using a squared exponential kernel. When all X_i in MCBO are modelled with squared exponential kernels and $\forall i$, $d_i = d, q_i = q$, we have a maximum information gain of $\gamma_T = \mathcal{O}(d(Kd+q)(\log T)^{Kd+q+1})$, resulting in a cumulative regret that is exponential in K and exponential in N. However, note that $m \geq K + N$. For several common graphs, the exponential scaling in N and K could be significantly more favorable than exponential scaling in m. Consider the case of $\mathcal{G}$ having the binary-tree-like structure in Figure 2 (binary tree), where $N = \log(m)$ and $K = 2$. In such settings, the cumulative regret of MCBO will have only polynomial dependence on m.

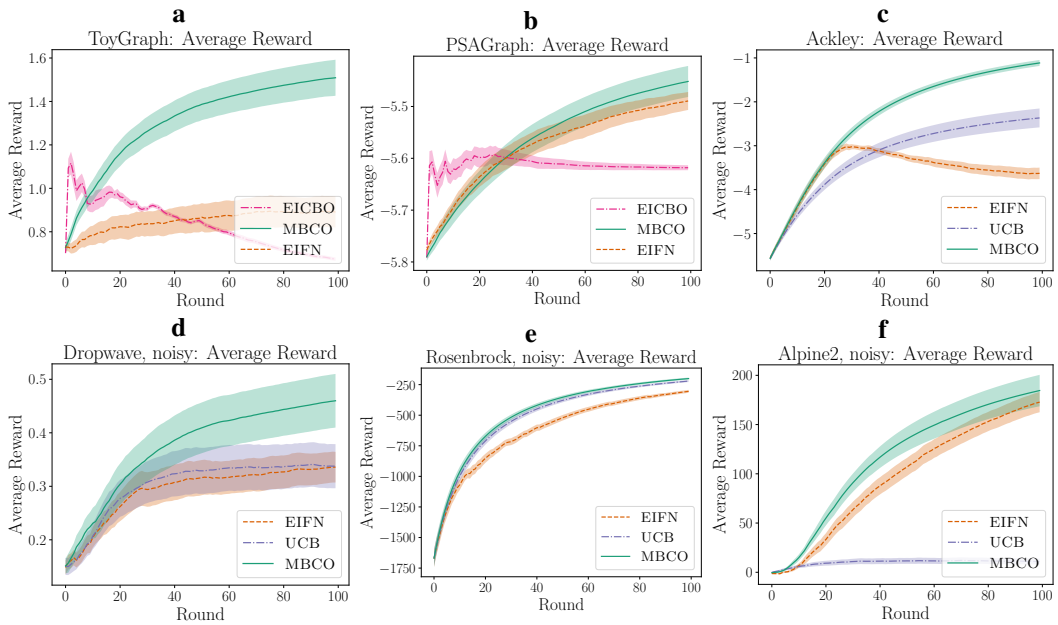

Figure 3: **(a, b)** EICBO does not achieve monotonically improving average regret on CBO tasks, while MCBO achieves high average reward on both tasks. **(c)** On the noiseless Ackley function networks task, EIFN also does not achieve monotonically improving average reward. **(d, e, f)** On a range of noisy function networks tasks, MCBO is at as good as and often higher performing than EIFN and UCB baselines.

5 Experiments

In this section, we empirically validate MCBO on six problems taken from previous CBO or function network works (Aglietti et al., 2020b; Astudillo & Frazier, 2021). The DAGs corresponding to each task are given in Fig. 2. We provide an open-source implementation of MCBO[4].

Baselines We compare our approach with three baselines (i) Expected Improvement for Function Networks (EIFN) (Astudillo & Frazier, 2021); (ii) Causal Bayesian Optimization (EICBO) (Aglietti et al., 2020b); and (iii) standard upper confidence bound (UCB) (Srinivas et al., 2010) which models the objective given interventions with a single GP (see Fig. 1 a). Importantly, EICBO cannot be applied in all the settings considered below because it only models specifically hard interventions.

Experimental setup We report the average reward as a function of the number of system interventions performed. The average reward at time T is defined by $\sum_{t=0}^{T} \mathbb{E}_{\omega}[Y \mid \boldsymbol{a}_t]$ and is directly inversely related to cumulative regret in that high average reward is equivalent to low cumulative regret. This matches the performance metric studied in our analysis. Average reward and cumulative regret are natural metrics for many real applications, like crop management, in which we want consistently high yield, and the healthcare-inspired setting we study in the experiments, in which we seek good treatment for more than a single patient. In the appendix, we show experiments measuring simple regret. We report mean performance over 20 random seeds, with error bars showing $\pm \sigma/\sqrt{20}$ where σ is the standard deviation across the repeats. All figures that are referenced but not in the main paper can be found in the appendix.

To implement MCBO, we use a Monte Carlo estimate to compute the expectation over $\boldsymbol{\omega}$ in Eq. (11). For the guarantees in Theorem 1 to hold, $\{\beta_t\}_{t=0}^{T}$ must be chosen so that the model is calibrated at all time-steps as in Eq. (9). In practice, we select a single β such that $\beta_t = \beta, \forall t$. Choosing β too pessimistically will result in high regret, as demonstrated by the dependence of the guarantee on β^N. For UCB and MCBO, β is chosen by cross-validation as described in the appendix.

[4] `https://anonymous.4open.science/r/mcbo/`

Toy Experiment First, we evaluate on the synthetic ToyGraph setting from Aglietti et al. (2020b). ToyGraph is a hard intervention CBO task where $\mathcal{I} = \{\emptyset, \{0\}, \{1\}\}$. All methods start with 10 observational samples and 2 samples from different random interventions on each $I \in \mathcal{I}$. When EICBO obtains interventional data, it obtains noiseless samples because it is not designed for the noisy setting. By noiseless samples, we mean that for action $\boldsymbol{a}$ EICBO observes $\mathbb{E}_{\boldsymbol{\omega}}[Y \mid \boldsymbol{a}]$. Other methods obtain single samples from the distribution $Y, X \mid \boldsymbol{a}$.

In ToyGraph, $I = \{1\}$ is the optimal target, but to efficiently learn the optimal $a_{\{1\}}$ the agent must generalise from both observational data and interventional data on other targets $I = \{0\}$. EICBO models Y given a_I with a separate GP for every I and is consequently only able to make use of observational data and interventional data with the same target. Since MCBO learns a full system model, it incorporates all observations into the learned model, even when interventions do not match.

Figure 3 (a) shows that the average reward of MCBO (Algorithm 2) is favourable compared to the baselines. Both baselines are built on expected improvement, which will continue to explore even after high reward solutions are found. This explains the non-monotonic average reward of EICBO.

Healthcare Experiment PSAGraph is inspired by the DAG from a real healthcare setting (Ferro et al., 2015) and also benchmarked by Aglietti et al. (2020b). The agent intervenes by prescribing statins and/or aspirin, while specifying the dosage, to control prostate-specific antigen (PSA) levels. Here $\mathcal{I} = \{\emptyset, \{2\}, \{3\}, \{2, 3\}\}$ and all interventions are hard intervention. Initial sample sizes are the same as for ToyGraph. Figure 3 (b) again shows EICBO having nonmonotonic average reward and strong comparable performance of MCBO.

Noiseless Function Networks In addition to the hard intervention setting, we evaluate MCBO and the baselines on the four tasks from Astudillo & Frazier (2021). All systems have up to six nodes and varying graph structures. In function networks, actions can affect multiple system variables and system variables can be children of multiple actions. Function networks are deterministic, so $\boldsymbol{\omega} = \mathbf{0}$. MCBO (Algorithm 1) can be applied directly in this setting and the guarantees are also easily transferable. Like in Astudillo & Frazier (2021), there are no constraints (besides bounded domain) on actions, and the agent is initialized with $2A + 1$ samples from random actions, where A is the number of action nodes.

Figure 3 (c) and Figure 4 (a,b,c) show that MCBO achieves competitive average reward on all tasks. EIFN is better on Dropwave. Meanwhile, MCBO is substantially better than EIFN on the Ackley and Rosenbrock tasks. Overall, there are not sufficiently many tasks established in the literature to conclusively say which properties might make a task favor EIFN over MCBO. However, this would be interesting to understand in future work and likely relates to the wider conversation in BO comparing expected improvement and UCB algorithms (Merrill et al., 2021). We find that the naive UCB approach, which does not use the graph structure, generally performs poorly, especially on problems with larger graphs like Alpine2. On Ackley, EIFN does not achieve monotonically improving average reward. Since it is also based on EI and has strong performance in terms of best score, we suspect this can be explained by similar reasons as for EICBO.

Noisy Function Networks We modify three of the function networks settings to include an additive zero-mean Gaussian noise at every system variable, making $\boldsymbol{\omega}$ non-zero. EIFN is designed for deterministic function networks and has no convergence guarantee in this setting. Results in terms of average (Figure 3 d,e,f) and best reward (Figure 6 g, h, i) are comparable to the noiseless case, with MCBO and EIFN both performing well compared to UCB.

6 Conclusion

This paper introduces MCBO, a principled model-based approach to solving Bayesian optimization problems over structural causal models. Our approach explicitly models all variables in the system and propagates epistemic uncertainty through the model to select interventions based on the optimism principle. This allows MCBO to solve global optimization tasks in systems with known causal structure significantly more efficiently than prior work. We prove the first theoretical convergence guarantees for an algorithm solving the CBO problem and demonstrate that its theoretical advantages are reflected in strong empirical performance. Future work might consider how to apply the method to large graphs, where the sets of all possible discrete intervention targets cannot be efficiently enumerated.

REFERENCES

Virginia Aglietti, Theodoros Damoulas, Mauricio Álvarez, and Javier González. Multi-task Causal Learning with Gaussian Processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal Bayesian Optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020b.

Sami Alabed and Eiko Yoneki. BoGraph: Structured Bayesian Optimization From Logs for Expensive Systems with Many Parameters. In *Proceedings of the 2nd European Workshop on Machine Learning and Systems*, 2022.

Raul Astudillo and Peter I. Frazier. Bayesian Optimization of Function Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Nicola Branchini, Virginia Aglietti, Neil Dhir, and Theodoros Damoulas. Causal entropy optimization. *arXiv:2208.10981*, 2022.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient Model-Based Reinforcement Learning through Optimistic Policy Search and Planning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

Ana Ferro, Francisco Pina, Milton Severo, Pedro Dias, Francisco Botelho, and Nuno Lunet. Use of statins and serum levels of prostate specific antigen. *Acta Urológica Portuguesa*, 32(2):71–77, 2015.

Shunya Kusakawa, Shion Takeno, Yu Inatsu, Kentaro Kutsukake, Shogo Iwazaki, Takashi Nakano, Toru Ujihara, Masayuki Karasuyama, and Ichiro Takeuchi. Bayesian Optimization for Cascade-type Multi-stage Processes. *arXiv preprint*, 2021.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *AAAI Conference on Artificial Intelligence*, 2019.

Erich Merrill, Alan Fern, Xiaoli Fern, and Nima Dolatnia. An empirical study of bayesian optimization: Acquisition versus partition. *Journal of Machine Learning Research*, 2021.

Jonas Močkus. On Bayesian methods for seeking the extremum. In *Optimization Techniques*, 1975.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of International Conference on Machine Learning (ICML))*, 2010.

A APPENDIX
MODEL-BASED CAUSAL BAYESIAN OPTIMIZATION

A.1 PROOFS FOR THEORETICAL ANALYSIS

As stated in the main text, we closely follow Curi et al. (2020), particularly the proofs in their Appendix D, where they prove similar guarantees for a model-based reinforcement learning problem. In contrast to Curi et al. (2020), which models RL transition dynamics with a single GP for all timesteps, MCBO uses independent GPs for modelling the functional relation $\{f_1, \cdots, f_m\}$ and uses the causal graph $\mathcal{G}$ to determine the input and output variables of these functions.

This section is organized as follows. We first recap the notation essential for the analysis in Appendix A.1.1. In Appendix A.1.2, we further discuss the notion of model complexity Γ_T introduced in for RL setting (Curi et al., 2020) and applicable beyond the GP setting discussed in the main paper. We then bound the cumulative regret in terms of Γ_T in Appendix A.1.3. Finally, Appendix A.1.4, we prove the main result of the paper Theorem 1 that makes use of the previous analysis.

A.1.1 NOTATION

In the proofs, we use the following notation. Dimensionalities $\mathcal{X}_i \subset \mathbb{R}^{d_i}$, $\mathcal{A}_i \subset \mathbb{R}^{q_i}$, $\mathcal{Z}_i \subset \mathbb{R}^{p_i}$ where $p_i = \sum_{j \in pa_i} d_j$. The reparametrization function $\eta_i \colon \mathcal{Z}_i \times \mathcal{A}_i \to [-1, 1]^{d_i}$, or $\eta_i \colon \mathbb{R}^{p_i} \times \mathbb{R}^{q_i} \to [-1, 1]^{d_i}$, and we further denote a vector $\boldsymbol{\eta}_i = \eta_i(\boldsymbol{z}_i, \boldsymbol{a}_i)$. We write mean function $\boldsymbol{\mu}_i \colon \mathcal{Z}_i \times \mathcal{A}_i \to \mathcal{X}_i$ and variance $\Sigma_i \colon \mathcal{Z}_i \times \mathcal{A}_i \to \mathcal{X}_i$, where we further use the same bold notation to emphasize it as a vector $\boldsymbol{\mu}_i(\boldsymbol{z}_i, \boldsymbol{a}_i)$ and $\boldsymbol{\Sigma}_i(\cdot) = \mathrm{Var}[\tilde{f}(\cdot)]$ is a diagonal matrix of size $d_i \times d_i$. $\boldsymbol{\sigma}_i(\cdot) = \mathrm{diag}(\boldsymbol{\Sigma}_i(\cdot)) \in \mathbb{R}^{d_i}$.

Data $\mathcal{D}_t = \{\boldsymbol{z}_{1:t}, \boldsymbol{a}_{1:t}, \boldsymbol{x}_{1:t}\}$, where each element includes the evaluation over the graph $\mathcal{G}$, e.g., $\boldsymbol{x}_{1:t} = \{\boldsymbol{x}_{i,1:t}\}_i^m$, $\boldsymbol{x}_{i,1:t} = [\boldsymbol{x}_{i,1}, \cdots, \boldsymbol{x}_{i,t}]^\top$, $\boldsymbol{x}_{i,t} = [x_{i,t}^1, \cdots, x_{i,t}^{p_i}]^\top$ and $\mathcal{D}_t^i = \{\boldsymbol{z}_{1:t}^i, \boldsymbol{a}_{1:t}^i, \boldsymbol{x}_{1:t}^i\}$.

For when i is a root node (no parents) the value of $\boldsymbol{\sigma}(\boldsymbol{z}_{i,t})$ is always 0 – this notation prevents needing to specify edge cases in sums.

For readability, we omit the time subscript t notation when it is clear that it refers to the t-th round. We also omit the usage of $\boldsymbol{a}_i$ where it is obvious, e.g., referring to $\sigma_{i,t-1}(\boldsymbol{z}_i)$ instead of $\sigma_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)$. The norm notation $\|\cdot\|$ refers to ℓ_2-norm if no additional notation is given.

A.1.2 MODEL COMPLEXITY

The number of samples needed to learn a good action is related to the number of samples needed to learn all GP models. We use this idea in our analysis and therefore introduce a quantification of the model complexity of our entire model class $\mathcal{M}_T$:

$$\Gamma_T = \max_{A \subset \{\mathcal{Z} \times \mathcal{A}\}^T} \sum_{t=1}^{T} \sum_{i=0}^{m} \left\| \boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i) \right\|^2. \tag{13}$$

This measure of model complexity closely relates to the *maximum information gain* γ_T commonly used in the standard regret guarantees of Bayesian optimization (Srinivas et al., 2010). The information capacity γ_T is the maximum information gain about the unknown f that can be obtained from a fixed set of its noisy evaluations $\boldsymbol{x}_{1:T}$ (see Eq. (34)). Later we show that Γ_T can be bounded by a sum of the information gains for all m GPs. It is worth noting that Equation 13 may be a loose notion of model complexity because it assumes we can independently choose every $\boldsymbol{z}_i$, but for many graphs, there could be overlap in the $\boldsymbol{z}_i$s (two nodes could have a shared parent).

A.1.3 ANALYSIS IN TERMS OF GENERAL MODEL COMPLEXITY Γ_T

Theorem 2. *Consider causal global optimization Eq.* (4) *over the given causal model $\mathcal{G}$ with the true stochastic functional relation Eq.* (2) *that satisfy Assumptions 1–3. Let $\{\boldsymbol{a}_{i,t} \in \mathcal{A}_i, \boldsymbol{z}_{i,t} \in \mathcal{Z}_i\}_{i,t>0}$ denote the set of actions chosen by* MCBO *and parents realizations of node i, respectively. Then, for all $T \geq 1$, with probability at least $1 - \delta$, the regret of Algorithm 1 is bounded by $R_T \leq \mathcal{O}\big(L_\sigma^N \beta_T^N K^N \sqrt{Tm\Gamma_T}\big)$.*

The main steps of the proof are as follows. In Lemma 1, we prove the existence of the reparametrization function. In Lemma 2, we bound the instantaneous regret at iteration t in terms of the expected difference in y for the same set of actions in the optimistic vs true dynamics. In Lemma 3 and Corollary 1 we bound the difference between these two sets of dynamics before concluding the proof.

Lemma 1. *Under Assumption 3, for any i and any sequence $[\boldsymbol{x}_{i,t}]_{t=0}^{T}$ generated by the true functional relation Eq.* (2) *for some chosen $\{\boldsymbol{a}_{i,t}\}_t$, there exists a function $\eta_i \colon \mathcal{Z}_i \times \mathcal{A}_i \to [-1,1]^{d_i}$ such that $\boldsymbol{x}_{i,t} = \tilde{\boldsymbol{x}}_{i,t}$ if $\forall i \;\; \boldsymbol{\omega}_i = \tilde{\boldsymbol{\omega}}_i$.*

Proof. By the calibrated model Assumption 3 for each $i = 0, \dots, m$, we have elementwise bound $|f_i(\boldsymbol{z}_i, \boldsymbol{a}_i) - \boldsymbol{\mu}_i(\boldsymbol{z}_i, \boldsymbol{a}_i)| \leq \beta_i \boldsymbol{\sigma_i}(\boldsymbol{z}_i, \boldsymbol{a}_i)$. Thus for each $\boldsymbol{z}_i, \boldsymbol{a}_i$ there exists a vector $\boldsymbol{\eta_i}$ with values in $[-1,1]^{d_i}$ such that $f_i(\boldsymbol{z}_i, \boldsymbol{a}_i) = \boldsymbol{\mu}_i(\boldsymbol{z}_i, \boldsymbol{a}_i) + \boldsymbol{\Sigma_i}(\boldsymbol{z}_i, \boldsymbol{a}_i)\boldsymbol{\eta_i}$. Let the function $\eta_i(\cdot)$ return this vector for each parent's input and action input, then the result follows. □

Lemma 2. *Under Assumption 3, with probability at least $(1-\delta)$ $\forall t \geq 0$ the instantaneous regret r_t is bounded by*

$$r_t = \mathbb{E}[y|f, \boldsymbol{a}^*] - \mathbb{E}[y|f, \boldsymbol{a}_t] \leq \mathbb{E}\Big[y|\tilde{f}_t, \boldsymbol{a}_t\Big] - \mathbb{E}[y|f, \boldsymbol{a}_t]. \tag{14}$$

Proof. The result follows directly from,

$$\mathbb{E}[y|f, \boldsymbol{a}^*] \leq \mathbb{E}\Big[y|\tilde{f}_t, \boldsymbol{a}_t\Big]. \tag{15}$$

This is true by the definition of $\tilde{f}_t, \boldsymbol{a}_t$ as the argmax of Eq. (12). □

Lemma 3. *Under Assumptions 1–3, let $\bar{L}_f = 1 + L_{\mathrm{f}} + 2\beta_t L_\sigma$. Then, for all iterations $t > 0$, any functions $\eta_i \colon \mathbb{R}^{p_i} \times \mathbb{R}^{q_i} \to [-1,1]^{d_i}$ for all i and any sequence of $\boldsymbol{\omega}_i$ with $\tilde{\boldsymbol{\omega}}_i = \boldsymbol{\omega}_i$, $\boldsymbol{a} \in \mathcal{A}$ we have*

$$\|\boldsymbol{x}_{m,t} - \tilde{\boldsymbol{x}}_{m,t}\| \leq 2\beta_t K^N \bar{L}_f^N \sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_i)\| \tag{16}$$

Proof. We prove by induction on i.

Base case. Consider the base case $i = 0$. Because the nodes are topologically ordered we will have $pa_0 = \{\}$. Its realization therefore depends only on the chosen action. Formally, we assume $\boldsymbol{z}_0 = \emptyset$, $\boldsymbol{x}_0 = f(\boldsymbol{z}_0, \boldsymbol{a}_0) + \boldsymbol{\omega}_0$ and $\tilde{\boldsymbol{x}}_0 = \tilde{f}_0(\tilde{\boldsymbol{z}}_0, \tilde{\boldsymbol{a}}_0) + \tilde{\boldsymbol{\omega}}_0$. Since $\boldsymbol{\omega}_0 = \tilde{\boldsymbol{\omega}}_0$,

$$\begin{aligned}
\|\boldsymbol{x}_{0,t} - \tilde{\boldsymbol{x}}_{0,t}\| &= \|f_0(\boldsymbol{z}_0, \boldsymbol{a}_0) + \boldsymbol{\omega}_0 - \boldsymbol{\mu}_{0,t-1}(\boldsymbol{z}_0, \boldsymbol{a}_0) - \beta_t \boldsymbol{\Sigma}_{0,t-1}(\boldsymbol{z}_0, \boldsymbol{a}_0)\eta_0(\boldsymbol{z}_0, \boldsymbol{a}_0) - \tilde{\boldsymbol{\omega}}_0\| \\
&\leq \|f_0(\boldsymbol{z}_0, \boldsymbol{a}_0) - \boldsymbol{\mu}_{0,t-1}(\boldsymbol{z}_0, \boldsymbol{a}_0)\| + \|\beta_t \boldsymbol{\Sigma}_{0,t-1}(\boldsymbol{z}_0, \boldsymbol{a}_0)\eta_0(\boldsymbol{z}_0, \boldsymbol{a}_0)\| \\
&\leq 2\beta_t \|\boldsymbol{\Sigma}_{0,t-1}(\boldsymbol{z}_0, \boldsymbol{a}_0)\|
\end{aligned}$$

In the following, we omit the dependence on the action $\boldsymbol{a}$, e.g., using $f_i(\boldsymbol{z}_i)$ instead of $f_i(\boldsymbol{z}_i, \boldsymbol{a}_i)$ since we assume the actions to be the same for $\boldsymbol{x}_{i,t}$ and $\tilde{\boldsymbol{x}}_{i,t}$.

Induction step. Now assuming that $\|\boldsymbol{x}_{i-1,t} - \tilde{\boldsymbol{x}}_{i-1,t}\| \leq 2\beta_t K^{N_{i-1}} \bar{L}_f^{N_{i-1}} \sum_{j=0}^{i-1} \|\boldsymbol{\sigma}_{j,t-1}(\boldsymbol{z}_{j,t})\|$ we prove a similar result for the ith node.

$$\begin{aligned}
\|\boldsymbol{x}_{i,t} - \tilde{\boldsymbol{x}}_{i,t}\| &\overset{1}{=} \|f_i(\boldsymbol{z}_i) + \boldsymbol{\omega}_i - \boldsymbol{\mu}_{i,t-1}(\tilde{\boldsymbol{z}}_i) - \beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\eta_i(\tilde{\boldsymbol{z}}_i) - \tilde{\boldsymbol{\omega}}_i\| \\
&\overset{2}{=} \|f_i(\boldsymbol{z}_i) - \boldsymbol{\mu}_{i,t-1}(\tilde{\boldsymbol{z}}_i) - \beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\eta_i(\tilde{\boldsymbol{z}}_i) + f_i(\tilde{\boldsymbol{z}}_i) - f_i(\tilde{\boldsymbol{z}}_i)\| \\
&\overset{3}{=} \|f_i(\tilde{\boldsymbol{z}}_i) - \boldsymbol{\mu}_{i,t-1}(\tilde{\boldsymbol{z}}_i) - \beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\eta_i(\tilde{\boldsymbol{z}}_i) + f_i(\boldsymbol{z}_i) - f_i(\tilde{\boldsymbol{z}}_i)\| \\
&\overset{4}{\leq} \|f_i(\tilde{\boldsymbol{z}}_i) - \boldsymbol{\mu}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\| + \|\beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\eta_i(\tilde{\boldsymbol{z}}_i)\| + \|f_i(\boldsymbol{z}_i) - f_i(\tilde{\boldsymbol{z}}_i)\| \\
&\overset{5}{\leq} \beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\| + \beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\| + L_f \|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\|
\end{aligned}$$

$$
\begin{aligned}
&\overset{6}{=} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i)\| + L_f \|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\| \\
&\overset{7}{=} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\boldsymbol{z}}_i) + \boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i) - \boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i)\| + L_f \|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\| \\
&\overset{8}{\leq} 2\beta_t \left(\|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i)\| + L_\sigma \|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\|\right) + L_f \|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\| \\
&\overset{9}{\leq} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i)\| + (1 + L_f + 2\beta_t L_\sigma) \|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\| \\
&\overset{10}{\leq} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i)\| + (1 + L_f + 2\beta_t L_\sigma) \sum_{j \in pa_i} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\| \\
&\overset{11}{\leq} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i)\| + (1 + L_f + 2\beta_t L_\sigma) \sum_{j \in pa_i} 2\beta_t K^{N_j} \underbrace{(1 + L_f + 2\beta_t L_\sigma)}_{=:\bar{L}_f}{}^{N_j} \sum_{h=0}^{j} \|\boldsymbol{\sigma}_{h,t-1}(\boldsymbol{z}_{h,t})\| \\
&\overset{12}{\leq} 2\beta_t K^{N_i} \bar{L}_f^{N_i} \sum_{j=0}^{i} \|\boldsymbol{\sigma}_{j,t-1}(\boldsymbol{z}_{j,t})\| \qquad (17)
\end{aligned}
$$

where (1) follows the dynamics Eqs. (2) and (11). In (2), we assume the noise to be equal and add and subtract the same term. In (3) and (4), we reorder terms and apply the triangle inequality. In (5) and (6), we rely on the calibrated uncertainty and Lipschitz dynamics, then collect terms and use diagonality of the matrix $\boldsymbol{\Sigma}_{i,t-1}(\cdot)$. In (7) and (8), we add and subtract the same term and use the Lipschitz continuity of $\boldsymbol{\sigma}_{i,t-1}$. Finally, in (9) we add 1 to ensure that when we later upper bound this term by taking the exponential of it. (10) applies the triangle inequality. (11) follows the inductive hypothesis, and (12) is due to the depth of at least one parent j being $N_j = N_i - 1$. □

Corollary 1. *Under the assumptions of Lemma 3, for any sequence of $\eta_i \in [-1, 1]^{d_i}$, $\boldsymbol{\theta} \in \mathcal{D}$, and $t \geq 1$ we have that*

$$
\mathbb{E}[y|f, \boldsymbol{a}_t] - \mathbb{E}\left[y|\tilde{f}_t, \boldsymbol{a}_t\right] \leq 2\beta_t K^N \bar{L}_f^N \mathbb{E}_{\omega=\tilde{\omega}} \left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})\|\right] \qquad (18)
$$

Proof. That follows from Lemma 3. Y is just the final state in the network so

$$
\begin{aligned}
\mathbb{E}[y|f, \boldsymbol{a}_t] - \mathbb{E}\left[y|\tilde{f}_t, \boldsymbol{a}_t\right] &= \mathbb{E}[\|\boldsymbol{x}_{m,t} - \tilde{\boldsymbol{x}}_{m,t}\|] \\
&\leq 2\beta_t K^N \bar{L}_f^N \mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})\|\right]
\end{aligned}
$$

□

Lemma 4. *Under Assumption 3, let $L_Y = 2\beta_t \bar{L}_f^N$. Then, with probability at least $(1 - \delta)$ it holds for all $t \geq 0$ that*

$$
r_t^2 \leq L_Y^2 K^{2N} m \mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})\|_2^2\right] \qquad (19)
$$

Proof.

$$
r_t \leq \mathbb{E}\left[y|\tilde{f}_t, \boldsymbol{a}_t\right] - \mathbb{E}[y|f, \boldsymbol{a}_t] \leq 2\beta_t K^N \bar{L}_f^N \mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t,\boldsymbol{a}_{i,t}})\|\right] \qquad (20)
$$

$$
\leq L_Y K^N \mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})\|\right] \qquad (21)
$$

$$r_t^2 \leq L_Y^2 K^{2N} \left(\mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})\| \right] \right)^2 \tag{22}$$

$$\leq L_Y^2 K^{2N} \mathbb{E}\left[\left(\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})\| \right)^2 \right] \tag{23}$$

$$\leq L_Y^2 K^{2N} m \mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})\|_2^2 \right] \tag{24}$$

Last two lines are Jensen's inequality. □

Now we bound cumulative regret R_T.

Lemma 5. *Under the assumption of Assumptions 1–3, with probability at least* $(1-\delta)$ *it holds for all* $t \geq 0$ *that*

$$R_T^2 \leq T L_Y^2 m \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})^2\|_2^2 \right] \tag{25}$$

Proof.

$$R_T^2 = \left(\sum_{t=1}^{T} r_t \right)^2 \overset{①}{\leq} T \sum_{t=1}^{T} r_t^2 \overset{②}{\leq} T L_Y^2 m \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})^2\|_2^2 \right], \tag{26}$$

□

where ① is due to Jensen's inequality and ② follows Lemma 4. Similar to Curi et al. (2020), this bound is dependent on the data observed by the iteration t, making it hard to interpret in a general case. To this end, we further provide the worst-case bound dependent on the model complexity I_T.

Lemma 6. *Under Assumptions 1–3, let* $\boldsymbol{z}_{i,t}$ *and* $\mathcal{Z}_{i,t-1} \subseteq \mathcal{Z}_{i,t}$*, and* $\boldsymbol{a}_{i,t} \in \mathcal{A}_i$ *for all* $i, t > 0$ *with compact sets* $\mathcal{X}_{i,t}$ *and* $\mathcal{A}_{i,t}$*. Then, with probability at least* $(1-\delta)$ *it holds for all* $t \geq 0$ *that*

$$R_T^2 \leq T L_Y^2 K^{2N} m \Gamma_T \tag{27}$$

where

$$\Gamma_T = \max_{\substack{A\colon\, A = \cup_i A_i, \\ \forall i\colon\, A_i \subset \mathcal{Z}_i \times \mathcal{A}_i,\, |A_i| = T}} \sum_{t=1}^{T} \sum_{i=0}^{m} \left\| \boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i) \right\|^2 \tag{28}$$

Proof. As a consequence of $\boldsymbol{z}_{i,t} \in \mathcal{Z}_i$ we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=0}^{m} \left\| \sigma_{i,t}(\boldsymbol{z}_{i,t}, \boldsymbol{a}_{i,t})^2 \right\|^2 \right] \leq \Gamma_T \tag{29}$$

and the result follows. □

That concludes the proof for Theorem 2.

A.1.4 Proof of Theorem 1

We can bound Γ_T in Theorem 2 to get a bound that depends on the specific GP model used for each f_i. We can show that for most commonly kernels MCBO achieves sublinear (in T) regret.

Theorem 1. *Consider the optimization problem in Eq.* (4) *with known causal graph $\mathcal{G}$ and unknown functional relations Eq.* (2) *that satisfy Assumptions 1–3. Let $\{\boldsymbol{a}_t \in \mathcal{A}, \boldsymbol{z}_t \in \mathcal{Z}\}_{t>0}$ denote the set of actions chosen by* MCBO *and the parent variable realizations, respectively. Then, provided $\mathcal{A}_i$ is compact $\forall i$, for all $T \geq 1$, with probability at least $1-\delta$, the cumulative regret of Algorithm 1 is bounded by*

$$R_T \leq \mathcal{O}\Big(L_f^N L_\sigma^N \beta_T^N K^N m \sqrt{T\,\gamma_T}\Big)\,.$$

Proof. ***Step 1.*** *On mutual information $I_T(\boldsymbol{x}_{i,1:T}, f_{i,1:T})$ and maximum information gain γ_T*

In the following we consider the information gain for the node i, i.e., for $\boldsymbol{x}_{i,1:T} \in \mathbb{R}^{d_i \times T}$ and $f_{i,1:T} = [f(\boldsymbol{z}_{i,1}, \boldsymbol{a}_{i,1}), \ldots, f(\boldsymbol{z}_{i,T}, \boldsymbol{a}_{i,T})]^\top$ evaluated at points $A_i = \{\boldsymbol{z}_{i,1:T}, \boldsymbol{a}_{i,1:T}\}$, $A_i \in \mathcal{Z}_i \times \mathcal{A}_i$. In this step, for simplicity, we partially will omit the i-notation in the derivation for the information gain, when it is clear. Mutual information $I_T(\boldsymbol{x}_{1:T}, f_{1:T})$ is then defined as:

$$I(\boldsymbol{x}_{1:T}, f_{1:T}) = H(\boldsymbol{x}_{1:T}) - H(\boldsymbol{x}_{1:T}|f_{1:T}),$$

where $H(\cdot)$ denotes entropy. For the fitting the GPs, the following modelling assumptions are used: $\boldsymbol{x}_t|f_t \sim \mathcal{N}(f_t(\boldsymbol{z}_t, \boldsymbol{a}_t), \boldsymbol{\sigma}^2\boldsymbol{I})$ and $\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1} \sim \mathcal{N}(\boldsymbol{\mu}_{t-1}(\boldsymbol{z}_t, \boldsymbol{a}_t), \boldsymbol{\sigma}^2\boldsymbol{I} + \boldsymbol{\Sigma}_{t-1}(\boldsymbol{z}_t, \boldsymbol{a}_t))$, where $\boldsymbol{I} \in \mathbb{R}^{d \times d}$, and

$$\boldsymbol{\mu}_{t-1}(\boldsymbol{z}_t, \boldsymbol{a}_t) = [\boldsymbol{\mu}_{t-1,d}(\boldsymbol{z}_t, \boldsymbol{a}_t, 1), \ldots, \boldsymbol{\mu}_{t-1,d}(\boldsymbol{z}_t, \boldsymbol{a}_t, d)]^\top,$$
$$\boldsymbol{\sigma}_{t-1}(\boldsymbol{z}_t, \boldsymbol{a}_t) = diag(\boldsymbol{\Sigma}_{t-1}(\boldsymbol{z}_t, \boldsymbol{a}_t)) = [\boldsymbol{\sigma}_{t-1,d}(\boldsymbol{z}_t, \boldsymbol{a}_t, 1), \ldots, \boldsymbol{\sigma}_{t-1,d}(\boldsymbol{z}_t, \boldsymbol{a}_t, d)]^\top.$$

We assume that the components $\{\boldsymbol{x}_{i,t,l}\}_{l=0}^{d_i}$ are independent of each other. From Srinivas et al. (2010) we know that the mutual information for the ith GP model is:

$$I(\boldsymbol{x}_{1:T}, f_{1:T}) = H(\boldsymbol{x}_{1:T}) - H(\boldsymbol{x}_{1:T}|f_{1:T}) = \frac{1}{2}\sum_{t=1}^{T}\sum_{l=1}^{d} \ln\left(1 + \frac{\boldsymbol{\sigma}_{t-1,d}^2(\boldsymbol{z}_t, \boldsymbol{a}_t, l)}{\boldsymbol{\sigma}^2}\right). \tag{30}$$

Per component:

$$I(\boldsymbol{x}_{i,1:T,l}, f_{i,1:T,l}) = \frac{1}{2}\sum_{t=1}^{T} \ln\left(1 + \frac{\boldsymbol{\sigma}_{t-1,d}^2(\boldsymbol{z}_t, \boldsymbol{a}_t, l)}{\boldsymbol{\sigma}^2}\right). \tag{31}$$

Then, by the definition of maximum information gain $\tilde{\gamma}_{i,T,l}$ per node i in the graph:

$$\tilde{\gamma}_{i,T,l} := \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} I(\boldsymbol{x}_{i,1:T,l}, f_{i,1:T,l}) = \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} \frac{1}{2}\sum_{t=1}^{T} \ln\left(1 + \frac{\boldsymbol{\sigma}_{t-1,d}^2(\boldsymbol{z}_t, \boldsymbol{a}_t, l)}{\boldsymbol{\sigma}^2}\right) \tag{32}$$

Accordingly, we can write the maximum information gain between $\boldsymbol{x}_{i,1:T}$ and $f_{i,1:T}$ as follows:

$$\tilde{\gamma}_{i,T} := \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} I(\boldsymbol{x}_{i,1:T}, f_{i,1:T}) = \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} \frac{1}{2}\sum_{t=1}^{T}\sum_{l=1}^{d} \ln\left(1 + \frac{\boldsymbol{\sigma}_{t-1,d}^2(\boldsymbol{z}_t, \boldsymbol{a}_t, l)}{\boldsymbol{\sigma}^2}\right) \tag{33}$$

$$= \sum_{l=1}^{d} \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} \frac{1}{2}\sum_{t=1}^{T} \ln\left(1 + \frac{\boldsymbol{\sigma}_{t-1,d}^2(\boldsymbol{z}_t, \boldsymbol{a}_t, l)}{\boldsymbol{\sigma}^2}\right) = \sum_{l=1}^{d} \tilde{\gamma}_{i,T,l} \tag{34}$$

The upper bounds on $\gamma_{i,T,l}$ are provided in Srinivas et al. (2010) for widely used kernels and scale sublinearly in T, e.g, for the linear kernel $\gamma_{i,T,l} = \mathcal{O}((p_i + q_i)\log T)$, and in case of the squared exponential kernel $\gamma_{i,T,l} = \mathcal{O}((p_i + q_i)(\log T)^{p_i+q_i+1})$. Therefore, the upper bounds for $\tilde{\gamma}_{i,T}$ scale accordingly taking dimentionality d_i into account, e.g., for linear kernel $\gamma_{i,T} = \mathcal{O}(d_i(p_i+q_i)\log T)$, and in case of squared exponential kernel $\gamma_{i,T} = \mathcal{O}(d_i(p_i + q_i)(\log T)^{p_i+q_i+1})$.

Step 2. *On the bound for model complexity* Γ_T.

Here we show how the model complexity Eq. (13)

$$\Gamma_T \leq \sum_{i=0}^{m} \frac{1}{\ln(1+\sigma_i^{-2})} \tilde{\gamma}_{i,T}. \tag{35}$$

For readability, in the following, we denote $\max_A(\cdot) \coloneqq \max_{\substack{A\colon A=\cup_i A_i\colon \\ \forall i\colon A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\})^T}} (\cdot)$.

$$\begin{aligned}
\Gamma_T &= \max_{\{\mathcal{Z}\times\mathcal{A}\}^T} \sum_{t=1}^{T}\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)\|_2^2 \\
&\overset{①}{\leq} \max_A \sum_{t=1}^{T}\sum_{i=0}^{m} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)\|_2^2 \overset{②}{\leq} \sum_{i=0}^{m} \max_A \sum_{t=1}^{T} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)\|_2^2 \\
&\overset{③}{\leq} \sum_{i=0}^{m} \max_{A_i} \sum_{t=1}^{T} \|\boldsymbol{\sigma}_{i,t-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)\|_2^2 \overset{④}{\leq} \sum_{i=0}^{m} \max_{A_i} \sum_{t=1}^{T}\sum_{l=1}^{d_i} \|\sigma_{i,(t-1),d}(\boldsymbol{z}_i, \boldsymbol{a}_i, l)\|_2^2 \\
&\overset{⑤}{\leq} \sum_{i=0}^{m} \frac{2}{\ln(1+\sigma_i^{-2})} \underbrace{\max_{A_i} \frac{1}{2} \sum_{t=1}^{T}\sum_{l=1}^{d_i} \ln\left(1 + \frac{\sigma_{i,(t-1),d}(\boldsymbol{z}_i, \boldsymbol{a}_i, l)}{\sigma_i^2}\right)}_{\text{maximum information gain } Eq.\ (34)} \\
&\overset{⑥}{=} \sum_{i=0}^{m} \frac{2}{\ln(1+\sigma_i^{-2})} \tilde{\gamma}_{i,T} \overset{⑦}{=} \mathcal{O}(m\gamma_T),
\end{aligned}$$

where ① bounds $\mathcal{A}$ and $\mathcal{Z}$ with a box. ② is from the max over a sum. ③ is due to the assumption of $\boldsymbol{x}_{i,t}$ being independent of A_j, $j \neq i$, conditioned on A_i. ④ is due to Jensen's inequality. ⑤ is due to the fact that for any $s^2 \in [0, \sigma^{-2}]$ we can bound $s^2 \leq \frac{\sigma^{-2}}{\ln(1+\sigma^{-2})} \ln(1+s^2)$ Srinivas et al. (2010). This also holds for $s^2(\cdot) \coloneqq \sigma_i^{-2}\sigma_{i,(t-1),d}^2(\cdot)$ since $\sigma_i^{-2}\sigma_{i,(t-1),d}^2(\cdot) \leq \sigma_i^{-2}k(\cdot,\cdot) \leq \sigma_i^{-2}$ because $k_i(\cdot,\cdot) < 1\ \forall i$ (bounded kernel assumption). ⑥ due to Eqs. (30) and (34). Finally, in ⑦ we define $\gamma_T \coloneqq \max_i \tilde{\gamma}_{i,T}$ being the maximum value of the maximum information gains over graph nodes.

Step 3. *On sublinearity w.r.t. T of maximum information gain* γ_T.

Finally, we show that the upper bounds on γ_T scale sublinearly in T. That follows from Step 1 where we show that $\tilde{\gamma}_{i,T}$ scale sublinearly for most of the popularly used kernels. Particularly, that leads to $\gamma_T = \mathcal{O}(d(p+q)\log T)$ to linear kernel and $\gamma_T = \mathcal{O}(d(p+q)(\log T)^{p+q+1})$ in case of squared exponential kernel, where p, q, d correspond to the node $i' = \arg\max_i \tilde{\gamma}_{i,T}$. Since γ_T scales sublinearly in T, so does the cumulative regret bound in Theorem 1.

□

A.2 Maximizing the Acquisition Function

As discussed, our theoretical results assume access to an oracle that can maximize Eq. (10). Here we discuss how we approximate this oracle in practice.

In noiseless settings, instead of parameterizing each η_i as a neural network, we can parameterize it as a scalar. This is because with no noise, the inputs to η_i ($\boldsymbol{z}_i$ and $\boldsymbol{a}_i$) given $\boldsymbol{a}$ are fixed. This keeps the space of parameters to optimize over small, meaning we can use an identical optimization procedure to that used in EIFN by Astudillo & Frazier (2021) which is also an out of the box optimizer in the BoTorch package Balandat et al. (2020).

For noisy settings where each η_i is a neural network, we use our own optimizer which is quite simple. For each initialization of η parameters and $\boldsymbol{a}$, we perform stochastic gradient descent. We

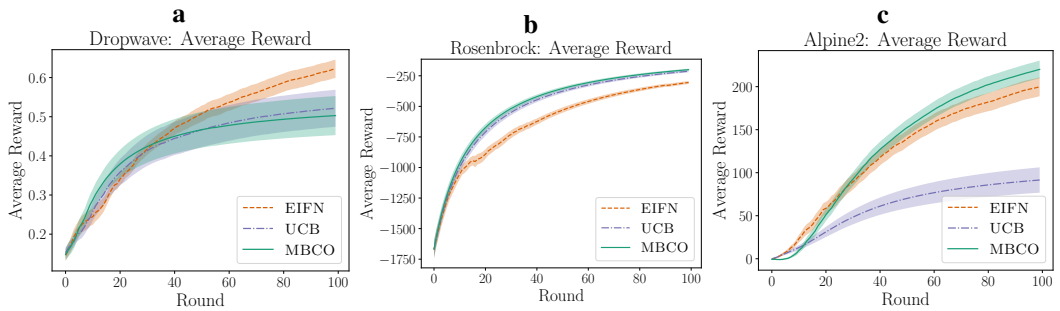

Figure 4: All average reward plots not used in the main paper. All settings are noiseless.

can do this because Eq. (12) is differentiable with respect to both $\boldsymbol{a}$ and the parameters of each η_i. After many repeated random initializations we will have many solution candidates. We select the one with the highest acquisition function value. We use a large number of different random initializations because the acquisition function may be very non-convex. Other approaches, such as those considered in Curi et al. (2020) could also be used to optimize our acquisition function.

On all noisy environments we estimate the expectation in the acquisitionn function using a Monte Carlo estimate with using 32 repeats for each gradient step. For noisy Dropwave we use 128 repeats because the environment is particularly noisy.

A.3 Experimental Setup

Here we give additional notes to explain the experimental setup and ensure full replicability.

The cross-validation for selecting β is performed across $\beta = \{0.05, 0.5, 5\}$. For a given performance metric and task, for example average reward on Dropwave, we select the β that on average performs best across all other tasks for the same performance metric, called β^*. We then report the results in terms of average reward for running the algorithm (UCB or MCBO) on Dropwave with $\beta = \beta^*$

For MCBO and EIFN, we use identical values for all shared hyperparameters (e.g. GP kernels) and use the original hyperparameters from Astudillo & Frazier (2021). CBO methods do not have hyperparameters comparable with these methods because CBO methods do not model individual system observations (see Section 2.1). We run the original EICBO source code effectively unmodified Aglietti et al. (2020b).

For CBO tasks, if observational data is selected, we give the agent, for any algorithm, a single observational sample. This is different to in Aglietti et al. (2020b) where 20 observational samples are given.

There is no noisy version of Ackley because adding noise can make the environment unstable by producing very large or very small rewards. For all environments and all X_i, the noisy environment adds unit-Gaussian noise, except Dropwave where we scale this noise by 0.1 for reasons of stability.

When applying MCBO to function networks we allow $\boldsymbol{a}_m$ to be nonzero, to match the function networks setting.

A.4 Further Experimental Analysis

Cumulative regret In Figure 4 we show the deterministic function networks plots not in the main paper. In Figure 5 we show the average reward for all β considered in the cross-validation.

Simple regret Here we also report best reward as a function of the number of rounds T. The best reward at time T is inversely related to simple regret and is defined by $\max_{\boldsymbol{a} \in \{\boldsymbol{a}_t\}_{t=0}^{T}} \mathbb{E}_{\omega}[Y \mid \boldsymbol{a}]$.

When performing the same cross-validation procedure as for the cumulative regret case, we find that the performance of MCBO varies drastically across task. This is shown in Figure 6. Selecting β is known to be a difficulty with UCB-based method (Merrill et al., 2021). Figure 7 shows the performance in terms of both average and best reward for all 3 possibilities for β. When selecting

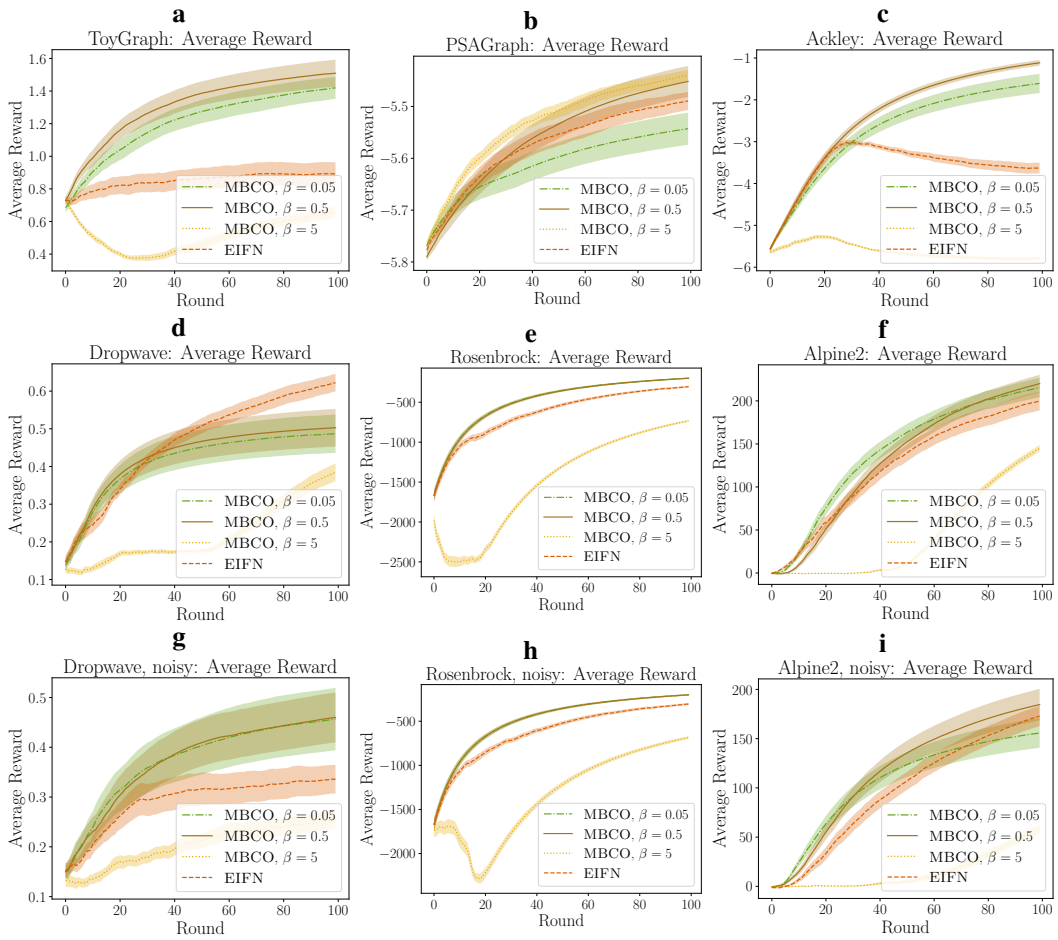

Figure 5: An ablation across β for the average reward across all tasks.

just the best $beta$ among the 3 tried, we find that MCBO is very strong on all tasks (with the exception of best score on ToyGraph, where it underperforms the baselines). The optimal β can vary by an order of magnitude or more across tasks in our case. This is likely because all of our settings have very different functional relationships and graph structures. As discussed in Section 4, this can be understood from our theoretical analysis.

The plots of average and best scores for other values of β (Figs. 5 and 7) suggest that β can be made larger for MCBO to trade-off exploration vs. exploitation depending on whether simple or cumulative regret is more of a priority. Meanwhile, expected improvement methods are focused almost almost entirely on exploration.

Runtimes In Figure 8 we show the total runtime of each method for all 100 rounds. EIFN and MCBO run on equivalent hardware, which has 4 times more ram than the hardware used for UCB and EICBO. MCBO generally has a much longer runtime than EIFN in noisy settings where ηs are parameterized by neural networks. In general in BO, because the unknown function is expensive to evaluate, we are less concerned about the time required to optimize the acquisition function. Our implementation for noisy settings could likely be sped-up with more parallelism to make it comparable to EIFN runtimes. In noiseless settings, where the optimization methods used are roughly equivalent between EIFN and MCBO, the two methods have comparable runtimes.

Non-monotonic regret of EICBO In Aglietti et al. (2020b) ToyGraph and PSAGraph are run with much larger initial sample sizes. We replicated the experiments with the same initial samples sizes as in Aglietti et al. (2020b) and find the same qualitative results (non-montonicity) for the average reward case. We use the smaller initial sample sizes in our paper to make the problem more difficult.

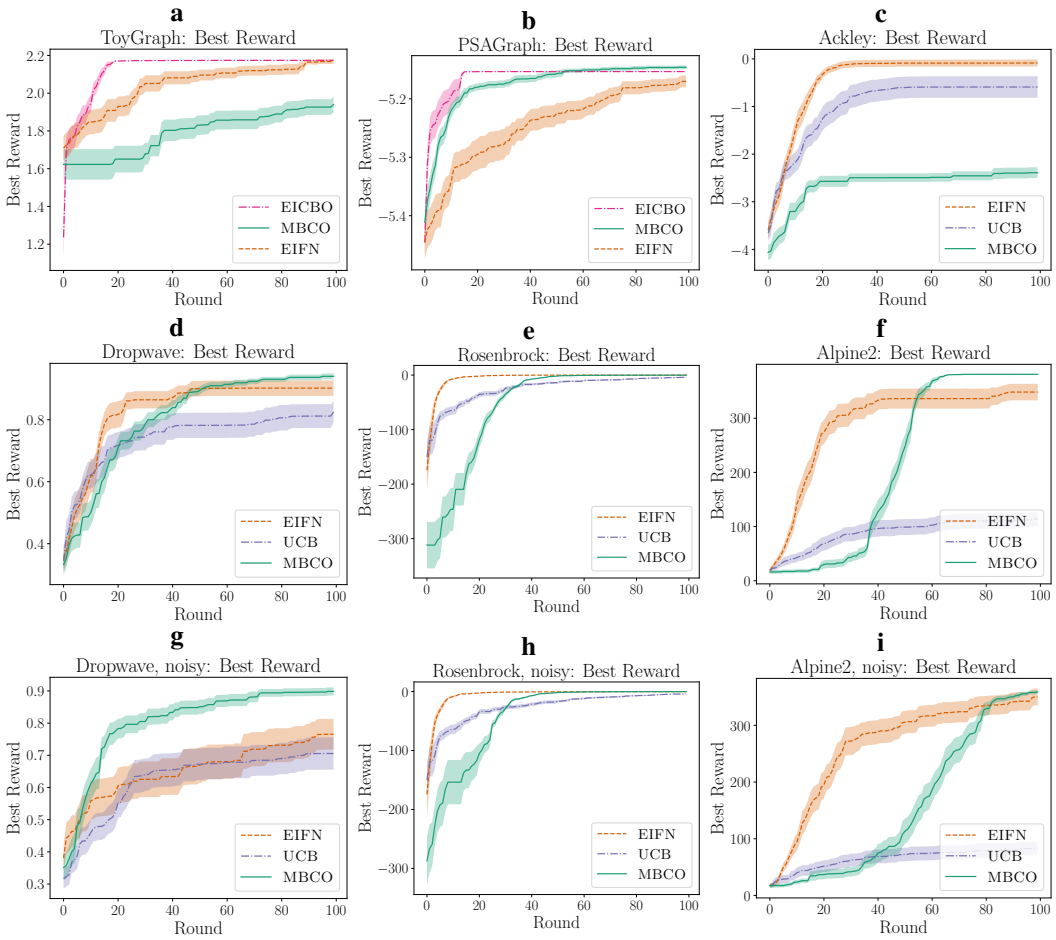

Figure 6: All best reward plots.

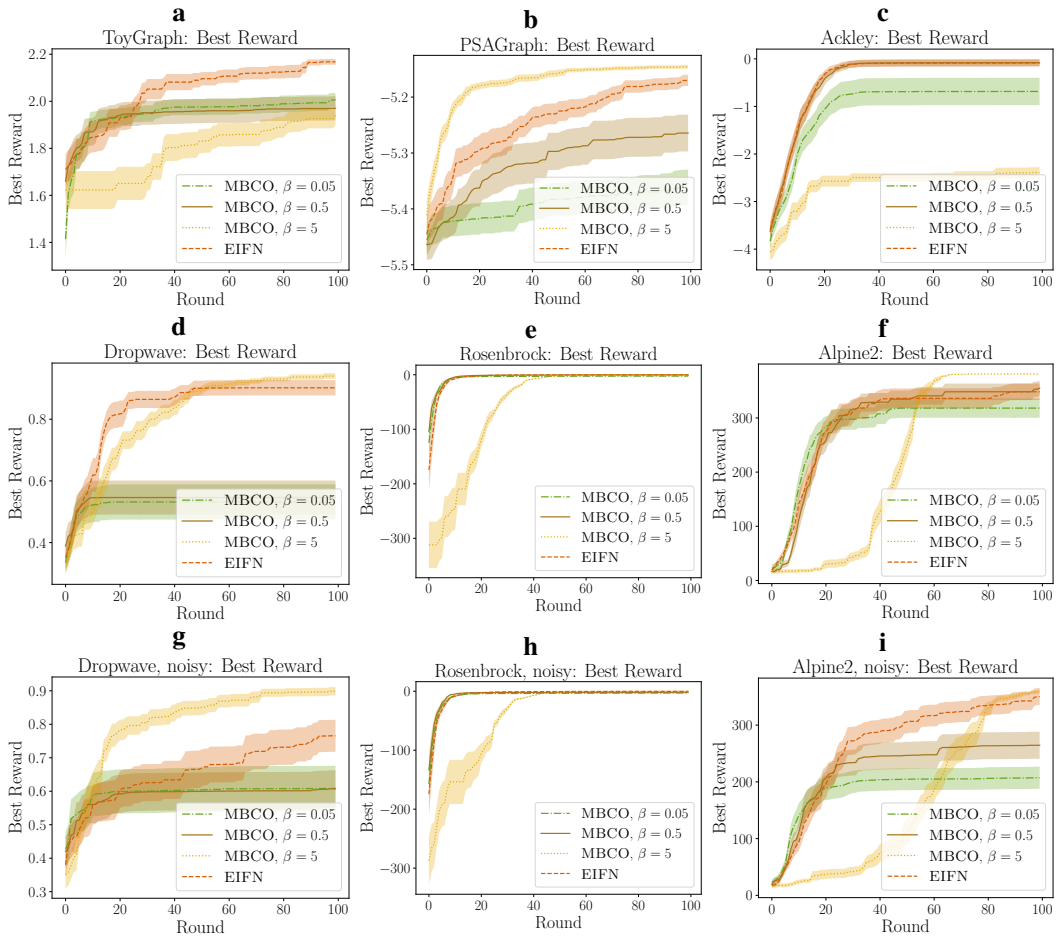

Figure 7: An ablation across β for the best reward across all tasks.

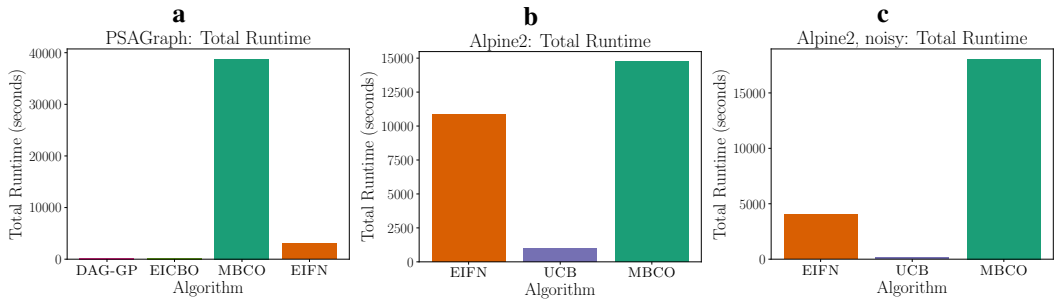

Figure 8: Runtimes of experiments for an **a**) CBO setting, **b**) noiseless function network setting, and **c**) noisy function network setting.