

# MODEL-BASED CAUSAL BAYESIAN OPTIMIZATION

**Scott Sussex**  
ETH Zürich  
scott.sussex@inf.ethz.ch

**Anastasiia Makarova**  
ETH Zürich

**Andreas Krause**  
ETH Zürich

## ABSTRACT

How should we intervene on an unknown structural causal model to maximize a downstream variable of interest? This optimization of the output of a system of interconnected variables, 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 *model-based causal Bayesian optimization* (MCBO), an algorithm 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. Empirically we find that MCBO compares favorably with existing state-of-the-art approaches.

## 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 (Moćkus, 1975). The key idea in BO is to quantify uncertainty in the unknown function via a probabilistic model, and then use this to navigate a trade-off between selecting inputs where the function output is favourable (exploitation) and selecting inputs to learn more about the function in areas of uncertainty (exploration). While most standard BO methods focus on a black-box setup (Figure 1 a), in practice, we often have more structure on the unknown function that can be used to improve sample efficiency.

In this paper, we *exploit structural knowledge in the form of a causal graph* specified by a directed acyclic graph (DAG). 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 action. This Causal BO setting has important potential applications, such as optimizing medical and ecological interventions (Aglietti et al., 2020b). For illustrative purposes, consider the 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 the causal graph structure for optimization, we propose *model-based causal Bayesian optimization* (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 uncertainty in all system components 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,

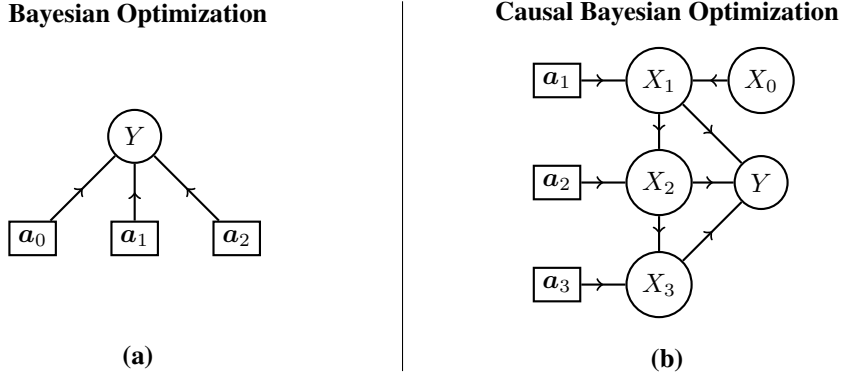


Figure 1: A visual comparison between the modelling assumptions of BO vs CBO. Circular nodes represent observed variables, squares represent action inputs and  $Y$  is the reward. Algorithms select  $\mathbf{a}$  before observing  $\mathbf{X}$  and  $Y$ . (a) In standard BO, the DAG has the structure shown regardless of the problem structure. (b) The DAG corresponding to our stylised agronomy example, where we aim to maximize crop yield  $Y$ . CBO takes this DAG as input for designing actions.  $X_0$  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. Each observation is modelled with its own Gaussian process.  $\mathbf{a}_1 \dots \mathbf{a}_3$  are possible interventions involving adding Nitrogen fertilizer to the soil.

MCBO achieves competitive performance on existing CBO benchmarks and a related setting called function network BO (Astudillo & Frazier, 2021b).

## Contributions

- We introduce MCBO, a model-based algorithm for causal Bayesian optimization that can be applied with very generic classes of interventions.
- Using MCBO we prove the first sublinear cumulative regret bound for CBO. We show how the bound scales depending on the graph structure. We demonstrate that CBO can lead to a potentially *exponential* improvement in cumulative regret, with respect to the number of actions, compared to standard BO.
- 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 CBO benchmarks and the related setting of function network BO. Our results show that MCBO performs favorably compared to methods designed specifically for these tasks.

## 2 BACKGROUND AND PROBLEM STATEMENT

We consider the problem of an agent interacting with an SCM for  $T$  rounds in order to maximize the value of a particular target variable. We start with introducing SCMs and the kinds of interventions an agent can perform on an SCM. In the following, we denote with  $[m]$  the set of integers  $\{0, \dots, m\}$ .

**Structural Causal Models** An SCM is described by a tuple  $\langle \mathcal{G}, Y, \mathbf{X}, \mathbf{F}, \boldsymbol{\Omega} \rangle$  of the following elements:  $\mathcal{G}$  is a known DAG;  $Y$  is the reward variable of interest;  $\mathbf{X} = \{X_i\}_{i=0}^{m-1}$  is a set of observed random variables; the set  $\mathbf{F} = \{f_i\}_{i=0}^m$  defines the functional relations between these variables; and  $\boldsymbol{\Omega} = \{\Omega_i\}_{i=0}^m$  is a set of independent noise variables with zero-mean and known distribution.

We use the notation  $Y$  and  $X_m$  interchangeably and assume the elements of  $\mathbf{X}$  are topologically ordered, i.e.,  $X_0$  is a root and  $X_m$  is a leaf. We use the notation  $pa_i \subset \{0, \dots, m\}$  for the indices of the parents of the  $i$ th node, and  $\mathbf{Z}_i = \{X_j\}_{j \in pa_i}$  for the parents of the  $i$ th node. We sometimes use  $X_i$  to refer to both the  $i$ th node and the  $i$ th random variable.

Each  $X_i$  is generated according to the function  $f_i : \mathcal{Z}_i \rightarrow \mathcal{X}_i$ , taking the parent nodes  $\mathcal{Z}_i$  of  $X_i$  as input:  $\mathbf{x}_i = f_i(\mathbf{z}_i) + \boldsymbol{\omega}_i$ , where lowercase denotes a realization of the corresponding random variable. The reward is a scalar  $x_m \in \mathbb{R}$ . An observation  $X_i$  is defined over a compact set  $\mathbf{x}_i \in \mathcal{X}_i \subset \mathbb{R}^d$ , and its parents are defined over  $\mathcal{Z}_i = \prod_{j \in \text{pa}_i} \mathcal{X}_j$  for  $i \in [m-1]$ .

**Interventions** At each interaction round, the agent performs an *intervention* on the SCM. In this work, we consider two types of intervention models.

We consider a soft intervention model (Eberhardt & Scheines, 2007) where interventions are parameterized by controllable *action variables*. Let  $\mathcal{A}_i \subset \mathbb{R}^q$  denote the compact space of an action  $\mathbf{a}_i$  and  $\mathcal{A}$  be the space of all actions  $\mathbf{a} = \{\mathbf{a}_i\}_{i=0}^m$ . We represent the actions as additional nodes in  $\mathcal{G}$  (see Fig. 1):  $\mathbf{a}_i$  is a parent of only  $X_i$ , and hence an additional input to  $f_i$ . Since  $f_i$  is unknown, in our soft intervention model, the agent does not know a priori the functional effect of  $\mathbf{a}_i$  on  $X_i$ . A simple example of a soft intervention is a shift intervention  $\mathbf{x}_i = f_i(\mathbf{z}_i, \mathbf{a}_i) + \boldsymbol{\omega}_i = g_i(\mathbf{z}_i) + \mathbf{a}_i + \boldsymbol{\omega}_i$  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.

While our theoretical results will focus on data obtained via soft interventions, our experiments also consider two other data sources. First, we consider a hard intervention model: hard interventions (often referred to as *do-interventions*) modify the targeted variable to a specific distribution independently of the variable’s parents. For example, a doctor sets the dosage of a patient’s medication, which fixes the dosage to a specific value (Aglietti et al., 2020b). Second, a special case under both intervention models is the collection of observational data, which is when no intervention is performed on the system. In the soft intervention model, not intervening on node  $i$  is equivalent to setting  $\mathbf{a}_i = \mathbf{0}$ . An example would be not applying any Nitrogen fertilizer to the soil. In practice, the agent may have access to some previous observational data before its first interaction with the system. In the following, we introduce the problem setup under the soft intervention model and then adapt it to the hard intervention model.

**Constraints on interventions** In many applications, we may not be able to intervene on all nodes simultaneously. For example, a farmer may only have the capacity to apply fertilizer at two out of three possible time windows in Fig. 1. This results in an action space with cardinality constraints, written as

$$\mathcal{A} = \left\{ \mathbf{a} = \{\mathbf{a}_i\}_{i=0}^m : \sum_{i=0}^m \mathbf{1}_{[\mathbf{a}_i \neq \mathbf{0}]} \leq c, c \geq 1 \right\}. \quad (1)$$

**Problem statement** We consider the problem of an agent sequentially interacting with an SCM, with known DAG  $\mathcal{G}$  and a fixed but unknown set of functions  $\mathbf{F} = \{f_i\}_{i=1}^m$  with  $f_i : \mathcal{Z}_i \times \mathcal{A}_i \rightarrow \mathcal{X}_i$ . At round  $t$  we select actions  $\mathbf{a}_{:,t} = \{\mathbf{a}_{i,t}\}_{i=0}^m$  and obtain observations  $\mathbf{x}_{:,t} = \{\mathbf{x}_{i,t}\}_{i=0}^m$ , where we add an additional subscript to denote the round of interaction. The action  $\mathbf{a}_{i,t}$  and the observation  $\mathbf{x}_{i,t}$  are related by

$$\mathbf{x}_{i,t} = f_i(\mathbf{z}_{i,t}, \mathbf{a}_{i,t}) + \boldsymbol{\omega}_{i,t}, \quad \forall i \in [m]. \quad (2)$$

If  $i$  corresponds to a root node, the parent vector  $\mathbf{z}_{i,t}$  denotes an empty vector, and the output of  $f_i$  only depends on the action  $\mathbf{a}_{i,t}$ . Since we cannot intervene on the target variable  $X_m$ , we fix  $\mathbf{a}_m = \mathbf{0}$ . The reward is given by

$$y_t = f_m(\mathbf{z}_{m,t}, \mathbf{a}_{m,t}) + \boldsymbol{\omega}_{m,t}, \quad (3)$$

which implicitly depends on the whole intervention vector  $\mathbf{a}_{:,t}$ . We define the action that maximizes the expected reward by

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in \mathcal{A}} \mathbb{E}[y|\mathbf{a}], \quad (4)$$

where, unless otherwise stated, expectations are taken over noise  $\boldsymbol{\omega}$ .

**Performance metric** Our agent’s goal is to design a sequence of interventions  $\{\mathbf{a}_{:,t}\}_{t=0}^T$  that achieves a high average expected reward. We hence study the *cumulative regret* (Lattimore & Szepesvári, 2020) over a time horizon  $T$ :

$$R_T = \sum_{t=1}^T \left[ \mathbb{E}[y|\mathbf{a}^*] - \mathbb{E}[y|\mathbf{a}_{:,t}] \right]. \quad (5)$$

A sublinear growth rate of  $R_T$  with  $T$  implies vanishing average regret:  $R_T/T \rightarrow 0$  as  $T \rightarrow \infty$ . As an alternative to cumulative regret, one can also study the simple regret  $\mathbb{E}[y|\mathbf{a}^*] - \mathbb{E}[y|\mathbf{a}_T]$ . The most appropriate metric depends on the application. In the Nitrogen fertilizer example, cumulative regret might be preferable because we care about obtaining high crop yields across all years, not just in one final year.

**Index notation** Let  $\mathbf{x}_{i,t} = [\mathbf{x}_{i,t,1}, \dots, \mathbf{x}_{i,t,d}]^T$  denote a vector where  $\mathbf{x}_{i,t,l}$  indicates indexing the component  $l \in [d]$  of the  $t$ th timepoint of the observations at the node  $i \in [m]$ . For functions with vector output, e.g.,  $\mathbf{f}_i : \mathcal{Z}_i \rightarrow \mathcal{X}_i$ , we sometimes consider notation with additional input to  $\mathbf{f}_i$  that indicates the output dimension:  $\mathbf{f}_i(\mathbf{z}, \mathbf{a}) = [f_i(\mathbf{z}, \mathbf{a}, 1), \dots, f_i(\mathbf{z}, \mathbf{a}, d)]^T$ .

**Regularity assumptions** We consider standard smoothness assumptions for the unknown functions  $\mathbf{f}_i : \mathcal{S} \rightarrow \mathcal{X}_i$  defined over a compact domain  $\mathcal{S}$  (Srinivas et al., 2010). In particular, for each node  $i \in [m]$ , we assume that  $\mathbf{f}_i(\cdot)$  belongs to a reproducing kernel Hilbert space (RKHS)  $\mathcal{H}_{k_i}$ , a space of smooth 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 kernel function  $k_i : \tilde{\mathcal{S}} \times \tilde{\mathcal{S}} \rightarrow \mathbb{R}$  where  $\tilde{\mathcal{S}} = \mathcal{S} \times [d]^1$ . We also assume that  $k_i(s, s') \leq 1$  for every  $s, s' \in \tilde{\mathcal{S}}^2$ . Moreover, the RKHS norm of  $\mathbf{f}_i(\cdot)$  is assumed to be bounded  $\|\mathbf{f}_i\|_{k_i} \leq \mathcal{B}_i$  for some fixed constant  $\mathcal{B}_i > 0$ . Finally, to ensure the compactness of the domains  $\mathcal{Z}_i$ , we assume that the noise  $\boldsymbol{\omega}$  is bounded, i.e.,  $\boldsymbol{\omega}_i \in [-1, 1]^d$ .<sup>3</sup>

**Problem statement under hard interventions** Under a hard intervention model, instead of selecting an action  $\mathbf{a}$  in Eq. (4), the agent must select both a set of intervention targets  $I \in \mathcal{I} \subset \mathcal{P}([m-1])$  and their values  $\mathbf{a}_I \in \mathcal{A}_I \in \mathcal{A}$ . For hard intervention we can rewrite Eq. (2) as

$$\mathbf{x}_i = \begin{cases} \mathbf{a}_i & \text{if } i \in I \\ \mathbf{f}_i(\mathbf{z}_i) + \boldsymbol{\omega}_i & \text{otherwise,} \end{cases} \quad \forall i \in [m], \quad (6)$$

where  $\mathbf{f}_i$  is unknown and employs the same regularity assumptions. Further constraints similar to Eq. (1) can be placed on either the intervention nodes  $I$  or action values  $\mathbf{a}_I$ . Finally, observational data corresponds to the empty intervention set  $I = \emptyset$ .

## 2.1 RELATED WORK

Optimal decision-making in SCMs has been the subject of several recent works, for example, in the bandit setting (Lattimore et al., 2016; Bareinboim et al., 2015). Aglietti et al. (2020b) introduce causal BO (CBO) focusing on hard interventions. CBO considers unobserved confounding and uses the do-calculus to estimate  $Y$  given  $I, \mathbf{a}_I$  using both observational data and interventional data with the same intervention targets  $I$ . Aglietti et al. (2020a) extend CBO to make use of data obtained from hard interventions with different intervention targets. While both methods use do-calculus to estimate causal effects, they do not learn the full system model. Branchini et al. (2022) and Alabed & Yoneki (2022) extend these works to explore the CBO setting in the case of unknown DAG.

Function network BO (FNBO) (Astudillo & Frazier, 2021b) is similar to the CBO setup with soft interventions and the proposed algorithm uses an expected improvement acquisition function to select actions. MCBO generalizes the FNBO setup to a richer class of problems since the causal model formalism allows for modelling hard interventions. Moreover, in contrast to MCBO, FNBO assumes the system is noiseless which might be a restrictive assumption in practice. Kusakawa et al. (2021) study stochastic function networks in the special case of a chain graph. Similar to MCBO, they develop a UCB-based acquisition function and provide an accompanying cumulative regret guarantee. However, they do not employ a reparameterization trick to show how to optimize the acquisition function. Kusakawa et al. (2021) do not perform any empirical study of the proposed UCB-based method, but evaluate expected improvement and credible interval methods

<sup>1</sup>For vector-valued functions coming from an RKHS we consider a scalar-valued function where the output index is part of the function input, as described in Curi et al. (2020) Appendix F.

<sup>2</sup>This is known as the bounded variance property, and it holds for many common kernels.

<sup>3</sup>This assumption can be relaxed to  $\boldsymbol{\omega}_i$  being 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  $\mathcal{Z}_i$ .

also developed for the chain graph setting. Both FNBO and CBO are part of a wider research direction on using intermediate observations from the computation of the unknown function to improve the sample efficiency of BO algorithms (Astudillo & Frazier, 2021a).

In concurrent work, Varici et al. (2022) consider a similar setup and prove a cumulative regret guarantee for a UCB-based algorithm with soft interventions. However, they focus only on linear SCMs and a binary action space for each intervention target, while MCBO applies to non-linear SCMs and a continuous action space.

The function class studied in this paper is similar to that of deep Gaussian processes (GPs) (Damianou & Lawrence, 2013) in that MCBO models  $Y$  as a composition of GPs given  $\mathbf{a}$ . 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 to design an algorithm for sample efficient reinforcement learning.

### 3 ALGORITHM

In this section, we propose the MCBO algorithm, describing the probabilistic model and acquisition function used. We first introduce MCBO under the soft intervention setup and then describe how to adapt it to hard interventions.

**Statistical model** We use Gaussian processes (GPs) for learning the RKHS functions  $f_0, \dots, f_m$  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. For all  $i \in [m]$ , let  $\boldsymbol{\mu}_{i,0}$  and  $\boldsymbol{\sigma}_{i,0}^2$  denote the prior mean and variance functions for each  $f_i$ , respectively. Since  $\boldsymbol{\omega}_i$  is bounded, it is also subgaussian and we denote the variance by  $\rho_i^2$ . The corresponding posterior GP mean and variance, denoted by  $\boldsymbol{\mu}_{i,t}$  and  $\boldsymbol{\sigma}_{i,t}^2$  respectively, are computed based on the previous evaluations  $\mathcal{D}_t = \{\mathbf{z}_{:,1:t}, \mathbf{a}_{:,1:t}, \mathbf{x}_{:,1:t}\}$ . In particular, for each function  $f_i(\cdot, \cdot, l)$  defined by the given kernel  $k_i$  and output component  $l$ :

$$\boldsymbol{\mu}_{i,t}(\mathbf{z}_i, \mathbf{a}_i, l) = \mathbf{k}_{i,t}(\mathbf{z}_i, \mathbf{a}_i, l)^\top (\mathbf{K}_t + \rho_i^2 \mathbf{I})^{-1} \text{vec}(\mathbf{x}_{i,1:t}), \quad (7)$$

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

where  $\mathbf{I}$  denotes the identity matrix,  $\text{vec}(\mathbf{x}_{i,1:t}) = [\mathbf{x}_{i,1,1}, \mathbf{x}_{i,1,2}, \dots, \mathbf{x}_{i,t,d}]^\top$  and for  $(t_1, l), (t_2, l') \in [(1, 1), (1, 2), \dots, (t, d)]$ :

$$\begin{aligned} [\mathbf{K}_t]_{(t_1, l), (t_2, l')} &= k_i((\mathbf{z}_{i,t_1, l}, \mathbf{a}_{i,t_1, l}, l); (\mathbf{z}_{i,t_2, l'}, \mathbf{a}_{i,t_2, l'}, l')), \\ \mathbf{k}_{i,t}(\mathbf{z}_i, \mathbf{a}_i, l)^\top &= [k_i((\mathbf{z}_{i,1,1}, \mathbf{a}_{i,1,1}, 1); (\mathbf{z}_i, \mathbf{a}_i, l)), \dots, k_i((\mathbf{z}_{i,t,d}, \mathbf{a}_{i,t,d}, d); (\mathbf{z}_i, \mathbf{a}_i, l))]^\top. \end{aligned}$$

We write  $\boldsymbol{\mu}_{i,t} = [\mu_{i,t}(\cdot, 1), \dots, \mu_{i,t}(\cdot, d)]^\top$  and similarly for  $\boldsymbol{\sigma}_{i,t}$ . We give more background on the posterior updates of vector-valued GPs in Appendix A.1.

At time  $t$ , the known set  $\mathcal{M}_t$  of statistically plausible functions  $\tilde{\mathbf{F}} = \{\tilde{f}_i\}_{i=0}^m$  (functions that lie inside the confidence interval given by the posterior of each GP) is defined as:

$$\begin{aligned} \mathcal{M}_t = \left\{ \tilde{\mathbf{F}} = \{\tilde{f}_i\}_{i=0}^m, \text{ s.t. } \forall i : \tilde{f}_i \in \mathcal{H}_{k_i}, \|\tilde{f}_i\|_{k_i} \leq \mathcal{B}_i, \right. \\ \left. \text{and } \left| \tilde{f}_i(\mathbf{z}_i, \mathbf{a}_i) - \boldsymbol{\mu}_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i) \right| \leq \beta_{i,t} \boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i), \forall \mathbf{z}_i \in \mathcal{Z}_i, \mathbf{a}_i \in \mathcal{A}_i \right\}. \quad (9) \end{aligned}$$

Here,  $\beta_{i,t}$  is a parameter that ensures the validity of the confidence bounds. Some examples of concentration inequalities under similar regularity assumptions as well as explicit forms for  $\beta_{i,t}$  can be found in Chowdhury & Gopalan (2019) and Srinivas et al. (2010). In the following, we set  $\beta_{i,t} = \beta_t$  for all  $i$  such that the confidence bounds in Eq. (9) are still valid.

**Algorithm 1** Model-based Causal BO (MCBO)

---

**Require:** Parameters  $\{\beta_t\}_{t \geq 1}, \mathcal{G}, \Omega$ , kernel functions  $k_i$  and prior means  $\boldsymbol{\mu}_{i,0} = \mathbf{0} \forall i \in [m]$

- 1: **for**  $t = 1, 2, \dots$  **do**
- 2:   Construct confidence bounds as in Eq. (9)
- 3:   Select  $\mathbf{a}_t \in \arg \max_{\mathbf{a} \in \mathcal{A}} \max_{\boldsymbol{\eta}(\cdot)} \mathbb{E}[y | \{\tilde{\mathbf{f}}\}, \mathbf{a}]$  as in Eq. (12)
- 4:   Observe samples  $\{\mathbf{z}_{i,t}, \mathbf{x}_{i,t}\}_{i=0}^m$
- 5:   Use  $\mathcal{D}_t$  to update posterior  $\{\boldsymbol{\mu}_{i,t}(\cdot), \boldsymbol{\sigma}_{i,t}^2(\cdot)\}_{i=0}^m$  as in Eqs. (7) and (8)
- 6: **end for**

---

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

---

**Require:** Parameters  $\{\beta_t\}_{t \geq 1}, \mathcal{G}, \Omega$ , kernel functions  $k_i$  and prior means  $\boldsymbol{\mu}_{i,0} = \mathbf{0} \forall i \in [m]$

- 1: **for**  $t = 1, 2, \dots$  **do**
- 2:   Construct confidence bounds as in Eq. (9)
- 3:   Select  $I, \mathbf{a}_I \in \arg \max_{I, \mathbf{a}_I} \max_{\boldsymbol{\eta}} \mathbb{E}[y | \{\tilde{\mathbf{f}}\}, do(X_I = \mathbf{a}_I)]$
- 4:   Observe samples  $\{\mathbf{z}_{i,t}, \mathbf{x}_{i,t}\}_{i=0}^m$
- 5:   Use  $\mathcal{D}_t$  to update posterior  $\{\boldsymbol{\mu}_{i,t}(\cdot), \boldsymbol{\sigma}_{i,t}^2(\cdot)\}_{i=0}^m$  as in Eqs. (7) and (8)
- 6: **end for**

---

**Acquisition function** At each round  $t$ , interventions are selected based on maximizing an acquisition function. Our acquisition function is based on the upper confidence bound acquisition function (Srinivas et al., 2010). That is, we optimistically pick interventions that yield the highest expected return among all system models that are still plausible given past observations:

$$\mathbf{a}_{:,t} = \arg \max_{\mathbf{a} \in \mathcal{A}} \max_{\tilde{\mathbf{F}} \in \mathcal{M}_t} \mathbb{E}_{\omega} [y | \tilde{\mathbf{F}}, \mathbf{a}]. \quad (10)$$

Note that Eq. (10) is not amenable to commonly used optimization techniques, due to the maximization over a set of functions with bounded RKHS norm. Therefore, following Curi et al. (2020), we use the reparameterization trick to write any  $\tilde{\mathbf{f}}_i \in \tilde{\mathbf{F}} \in \mathcal{M}_t$  using a function  $\boldsymbol{\eta}_i : \mathcal{Z}_i \times \mathcal{A}_i \rightarrow [-1, 1]^{d_i}$  as:

$$\tilde{\mathbf{f}}_{i,t}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) = \boldsymbol{\mu}_{i,t-1}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) + \beta_t \boldsymbol{\sigma}_{i,t-1}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) \circ \boldsymbol{\eta}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i), \quad (11)$$

where  $\tilde{\mathbf{x}}_i = \tilde{\mathbf{f}}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) + \tilde{\boldsymbol{\omega}}_i$  denotes observations from simulating actions in one of the plausible models, and not necessarily the true model.  $\circ$  denotes the elementwise multiplication of vectors. This reparameterization allows for rewriting our acquisition function in terms of  $\boldsymbol{\eta} : \mathcal{Z} \times \mathcal{A} \rightarrow [-1, 1]^{|\mathcal{X}|}$ :

$$\mathbf{a}_{:,t} = \arg \max_{\mathbf{a} \in \mathcal{A}} \max_{\boldsymbol{\eta}(\cdot)} \mathbb{E}_{\omega} [y | \tilde{\mathbf{F}}, \mathbf{a}], \quad \text{s.t. } \tilde{\mathbf{F}} = \{\tilde{\mathbf{f}}_{i,t}\} \text{ in Eq. (11)}. \quad (12)$$

Intuitively, the variables  $\boldsymbol{\eta}$  allow for choosing optimistic but plausible models given the confidence bounds. In practice, the function  $\boldsymbol{\eta}$  can be parameterized by, for example, a neural network, and then standard optimization techniques are applied. For the theory, we assume access to an oracle providing 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 heuristics that we use for approximating this oracle in Appendix A.3.

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(a), our algorithm and the associated guarantees reduce to standard BO (Srinivas et al., 2010).

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

## 4 THEORETICAL ANALYSIS

This section describes the convergence guarantees for MCBO under a soft intervention model, showing the first sublinear cumulative regret bounds for causal BO. We start by introducing additional technical assumptions required for the analysis.

**Assumption 1 (Functional relations).** All  $f_i \in \mathbf{F}$  are  $L_f$ -Lipschitz continuous.

**Assumption 2 (Continuity).**  $\forall i, t$ , the functions  $\mu_{i,t}, \sigma_{i,t}$  are  $L_\mu, L_\sigma$  Lipschitz continuous.

**Assumption 3 (Calibrated model).** All statistical models are calibrated w.r.t.  $\mathbf{F}$ , so that  $\forall i, t$  there exists a sequence  $\beta_t \in \mathbb{R}_{>0}$  such that, with probability at least  $(1 - \delta)$ , for all  $\mathbf{z}_i, \mathbf{a}_i \in \mathcal{Z}_i \times \mathcal{A}_i$  we have  $|\mathbf{f}_i(\mathbf{z}_i, \mathbf{a}_i) - \mu_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i)| \leq \beta_t \sigma_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i)$ , element-wise.

Assumption 1 follows directly from the regularity assumptions of Section 2. Assumption 2 holds if the RKHS of each  $f_i$  has a Lipschitz continuous kernel (see Curi et al. (2020), Appendix G). Assumption 2 restricts the convergence guarantees to soft interventions that affect their target variable in a smooth way, meaning that our analysis does not directly apply to the hard intervention model. 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  $i$ th GP prior uses the same kernel as the RKHS of  $f_i$  and that  $\beta_t$  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  $X_m$ :  $N = \max_i \text{dist}(X_i, X_m)$  where  $\text{dist}(\cdot, \cdot)$  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  $\mathcal{G}$ :  $K = \max_i |pa(i)|$ . The following theorem bounds the performance of MCBO in terms of cumulative regret.

**Theorem 1.** Consider the optimization problem in Eq. (4) with SCM satisfying Assumptions 1–3 where  $\mathcal{G}$  is known but  $\mathbf{F}$  is unknown. Then for all  $T \geq 1$ , with probability at least  $1 - \delta$ , the cumulative regret of Algorithm 1 is bounded by

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

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, such as linear and squared exponential kernels, resulting in an overall sublinear regret for MCBO. We refer to Appendix A.2.3 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 Fig. 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 modeled with squared exponential kernels that model output components independently, we have  $\gamma_T = \mathcal{O}(d(Kd + q)(\log T)^{Kd+q+1})$ , 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 the exponential scaling in  $m$ . Consider the case of  $\mathcal{G}$  having the binary-tree-like structure in appendix Fig. 3 (binary tree), where  $N = \log(m)$  and  $K = 2$ . In such settings, the cumulative regret of MCBO will have only *polynomial* dependence on  $m$ . We further discuss the bound in Theorem 1 for specific kernels in Appendix A.2.3 and discuss the dependence of  $\beta_T$  on  $T$  in Appendix A.2.4.

## 5 EXPERIMENTS

In this section, we empirically evaluate MCBO on six problems taken from previous CBO or function network papers (Aglietti et al., 2020b; Astudillo & Frazier, 2021b). The DAGs corresponding to each task are given in Fig. 3 of the appendix. We provide an open-source implementation of MCBO<sup>4</sup>.

<sup>4</sup><https://github.com/sseth/mcbo>

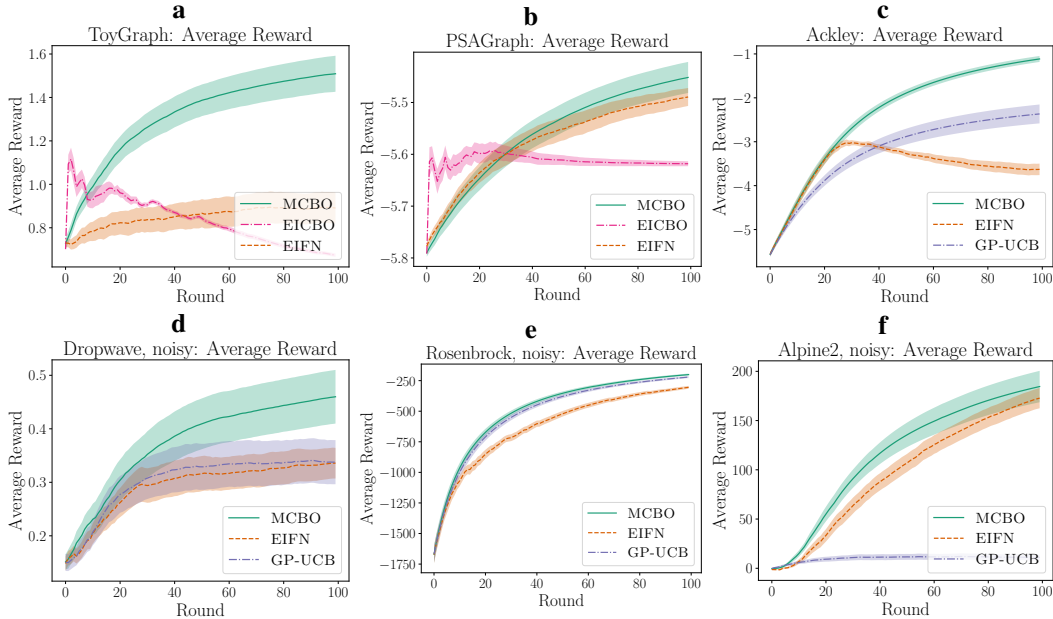


Figure 2: **(a, b)** EICBO does not achieve monotonically improving average reward on CBO tasks, while MCBO achieves high average reward on both tasks. **(c)** On the noiseless Ackley function networks task, EIFN also does not achieve monotonically improving average reward. **(d, e, f)** On a range of noisy function networks tasks, MCBO is at as good as and often higher performing than EIFN and GP-UCB baselines.

**Baselines** We compare our approach with three baselines (i) Expected Improvement for Function Networks (EIFN) (Astudillo & Frazier, 2021b); (ii) Causal Bayesian Optimization (EICBO) (Aglietti et al., 2020b); and (iii) standard upper confidence bound (GP-UCB) (Srinivas et al., 2010) which models the objective given interventions with a single GP (see Fig. 1 a). To enable a fair comparison, we only apply EICBO to the hard intervention setting, since it was designed specifically for a hard intervention model.

**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  $\sum_{t=0}^T \mathbb{E}_\omega[Y | \mathbf{a}_t]$  and is directly inversely related to cumulative regret in that high average expected reward is equivalent to low cumulative regret. This matches the performance metric studied in our analysis. Average expected 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 these experiments, where we seek good treatment outcomes for more than a single patient. In the appendix, we show experiments measuring the best expected reward of any action previously chosen, which is more similar to an inverse of the simple regret. We report mean performance over 20 random seeds, with error bars showing  $\pm \sigma/\sqrt{20}$  where  $\sigma$  is the standard deviation across the repeats. All figures that are referenced but not in the main paper can be found in the appendix.

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  $\beta$  such that  $\beta_t = \beta, \forall t$ . Choosing  $\beta$  too pessimistically will result in high regret, as demonstrated by the dependence of the guarantee on  $\beta^N$ . For GP-UCB and MCBO,  $\beta$  is chosen by cross-validation across tasks, as described in the appendix.

**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 samples from 2 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  $\mathbf{a}$  EICBO observes  $\mathbb{E}_\omega[Y | \mathbf{a}]$ . Other methods obtain single samples from the distribution  $Y, X | \mathbf{a}$ .



In ToyGraph,  $I = \{1\}$  is the optimal target, but to efficiently learn the optimal  $a_{\{1\}}$ , the agent must generalize 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 2 (a) shows that the average reward of MCBO (Algorithm 2) is favorable 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 interventions. Initial sample sizes are the same as for ToyGraph. Figure 2 (b) again shows EICBO having a 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 four tasks from Astudillo & Frazier (2021b). 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  $\omega = \mathbf{0}$ . MCBO (Algorithm 1) can be applied directly in this setting, and the guarantees are also easily transferable. Like in Astudillo & Frazier (2021b), 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 2 (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 GP-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, which is not unexpected given that it is based upon expected improvement.

**Noisy Function Networks** We modify three of the function networks settings to include an additive zero-mean Gaussian noise at every system variable, making  $\omega$  non-zero. EIFN is designed for deterministic function networks and has no convergence guarantees in this setting. Results in terms of average (Figure 2 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 GP-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 that have known causal structure with improved sample efficiency compared to prior works. We prove the first non-asymptotic convergence guarantees for an algorithm solving the causal Bayesian optimization 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.

## ACKNOWLEDGMENTS AND DISCLOSURE OF FUNDING

We thank Lars Lorch, Parnian Kassraie and Inura Usmanova for their feedback and anonymous reviewers for their helpful comments. This research was supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40 180545, and by the European Research Council (ERC) under the European Union’s Horizon grant 815943.

## 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. Thinking inside the box: A tutorial on grey-box bayesian optimization. In *2021 Winter Simulation Conference (WSC)*. IEEE, 2021a.
- Raul Astudillo and Peter I. Frazier. Bayesian Optimization of Function Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.
- 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.
- Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
- Nicola Branchini, Virginia Aglietti, Neil Dhir, and Theodoros Damoulas. Causal entropy optimization. *arXiv:2208.10981*, 2022.
- Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.
- 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.
- Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.
- 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.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- 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 Moćkus. 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.

Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *arXiv preprint arXiv:2208.12764*, 2022.

## A APPENDIX

### MODEL-BASED CAUSAL BAYESIAN OPTIMIZATION

#### A.1 BACKGROUND ON GPs

Here we give more background on the vector-valued GP models we use. We will drop the  $i$ -node index and consider the modelling of a single function with scalar output  $f : \mathcal{A} \times \mathcal{Z} \rightarrow \mathbb{R}$  in RKHS  $\mathcal{H}_k$  with kernel  $k : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  where  $\mathcal{S} = (\mathcal{Z} \times \mathcal{A})$ . We assume measurement noise with variance  $\rho^2$ . Assuming prior mean and variance  $\mu_0, \sigma_0$  and dataset  $\mathcal{D}_t = \{\mathbf{z}_{1:t}, \mathbf{a}_{1:t}, \mathbf{x}_{1:t}\}$  where  $\mathbf{x}_t \in \mathbb{R}$ , we get a posterior GP mean and variance given by updates

$$\mu_t(\mathbf{z}, \mathbf{a}) = \mathbf{k}_t(\mathbf{z}, \mathbf{a})^\top (\mathbf{K}_t + \rho^2 \mathbf{I})^{-1} \mathbf{x}_{1:t}, \quad (13)$$

$$\sigma_t^2(\mathbf{z}, \mathbf{a}) = k((\mathbf{z}, \mathbf{a}); (\mathbf{z}, \mathbf{a})) - \mathbf{k}_t(\mathbf{z}, \mathbf{a})^\top (\mathbf{K}_t + \rho^2 \mathbf{I})^{-1} \mathbf{k}_t(\mathbf{z}, \mathbf{a}), \quad (14)$$

where

$$\begin{aligned} (\mathbf{K}_t)_{t_1, t_2} &= k((\mathbf{z}_{t_1}, \mathbf{a}_{t_1}); (\mathbf{z}_{t_2}, \mathbf{a}_{t_2})), \\ \mathbf{k}_t(\mathbf{z}, \mathbf{a}) &= [k((\mathbf{z}_1, \mathbf{a}_1); (\mathbf{z}, \mathbf{a})) \dots, k((\mathbf{z}_t, \mathbf{a}_t); (\mathbf{z}, \mathbf{a}))], \end{aligned}$$

and  $\mathbf{I}$  is the identity matrix. This follows the standard GP update given in Rasmussen (2003).

In this work, we model functions with vector-valued outputs  $\mathbf{f} : \mathcal{A} \times \mathcal{Z} \rightarrow \mathcal{X} \in \mathbb{R}^d$  (we will continue to drop the  $i$  index). For this we follow Chowdhury & Gopalan (2019) and use a scalar-output GP that takes the component of the output vector under consideration as an input. That is, we model  $f(\cdot, l)$  (the  $l$ th output component of the function under study) where  $\mathbf{f} = [f(\cdot, 1), \dots, f(\cdot, d)]$ . We again assume that  $f$  is from RKHS  $\mathcal{H}_k$  but with kernel  $k : \tilde{\mathcal{S}} \times \tilde{\mathcal{S}} \rightarrow \mathbb{R}$  where  $\tilde{\mathcal{S}} = (\mathcal{Z} \times \mathcal{A} \times [d])$ . Assuming prior mean and variance  $\mu_0, \sigma_0$  and dataset  $\mathcal{D}_t = \{\mathbf{z}_{1:t}, \mathbf{a}_{1:t}, \mathbf{x}_{1:t}\}$  where  $\mathbf{x}_t \in \mathbb{R}^d$ , we get a posterior GP mean and variance given by updates

$$\mu_t(\mathbf{z}, \mathbf{a}, l) = \mathbf{k}_t(\mathbf{z}, \mathbf{a}, l)^\top (\mathbf{K}_t + \rho^2 \mathbf{I})^{-1} \text{vec}(\mathbf{x}_{1:t}), \quad (15)$$

$$\sigma_t^2(\mathbf{z}, \mathbf{a}, l) = k((\mathbf{z}, \mathbf{a}, l); (\mathbf{z}, \mathbf{a}, l)) - \mathbf{k}_t(\mathbf{z}, \mathbf{a}, l)^\top (\mathbf{K}_t + \rho^2 \mathbf{I})^{-1} \mathbf{k}_t(\mathbf{z}, \mathbf{a}, l), \quad (16)$$

where  $\text{vec}(\mathbf{x}_{1:t}) = [\mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \dots, \mathbf{x}_{t,d}]^\top$  and for  $(t_1, l), (t_2, l') \in [(1, 1), (1, 2), \dots, (t, d)]$ :

$$\begin{aligned} [\mathbf{K}_t]_{(t_1, l), (t_2, l')} &= k((\mathbf{z}_{t_1, l}, \mathbf{a}_{t_1, l}, l); (\mathbf{z}_{t_2, l'}, \mathbf{a}_{t_2, l'}, l')), \\ \mathbf{k}_t(\mathbf{z}, \mathbf{a}, l)^\top &= [k((\mathbf{z}_{1,1}, \mathbf{a}_{1,1}, 1); (\mathbf{z}, \mathbf{a}, l)), \dots, k((\mathbf{z}_{t,d}, \mathbf{a}_{t,d}, d); (\mathbf{z}, \mathbf{a}, l))]^\top. \end{aligned}$$

This follows the posterior update of Chowdhury & Gopalan (2019). The key idea is to use a single scalar-output GP with kernel  $k$  for modeling all output components, but introduce the component index as part of the input space. This requires introducing the notation  $\text{vec}(\mathbf{x}_{1:t})$  and ordering all observations, of all time points and all components, in a single index, since the GP update for component  $l$  considers observations of all other components. Under this GP model the components of  $\mathbf{f}$  need not be independent if kernel  $k$  is designed to model dependency between the components. In our work, we use one of these vector-valued GP models for each individual random variable in our causal model, leading to the reintroduction of the additional  $i$  index in Eqs. (7) and (8).

#### A.2 PROOFS FOR THE THEORETICAL ANALYSIS

Our analysis closely follows 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 modeling the functional relation  $\{f_1, \dots, f_m\}$  and uses the causal graph  $\mathcal{G}$  to determine the input and output variables of these functions.

This section is organized as follows. In Appendix A.2.1, we discuss a notion of model complexity  $\Gamma_T$  similar to the one introduced in the RL setting by Curi et al. (2020). We then bound the cumulative regret in terms of  $\Gamma_T$  in Appendix A.2.2. Finally, in Appendix A.2.3, we prove Theorem 1 by connecting our notion of model complexity with the maximum information gain of a GP model.

The norm notation  $\|\cdot\|$  refers to  $\ell_2$ -norm if no additional notation is given. We let  $\{\mathbf{a}_{i,t} \in \mathcal{A}_i, \mathbf{z}_{i,t} \in \mathcal{Z}_i\}_{i,t>0}$  denote the set of actions chosen by MCBO and the realizations of the parents of node  $i$ , respectively.

### A.2.1 MODEL COMPLEXITY

The number of samples needed to learn a low-regret action is related to the number of samples needed to learn all GP models in our SCM. This is analogous to the classic BO setting (Srinivas et al., 2010). We quantify the model complexity of our entire model class  $\mathcal{M}_T$  as

$$\Gamma_T = \max_{(\mathbf{z}, \mathbf{a}) \in \mathcal{A} \subset \{\mathcal{Z} \times \mathcal{A}\}^T} \sum_{t=1}^T \sum_{i=0}^m \|\boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\|^2. \quad (17)$$

This measure of model complexity closely relates to the *maximum information gain*  $\gamma_T$  used for proving regret guarantees in BO (Srinivas et al., 2010). For a single GP model,  $\gamma_T$  is the maximum information gain about the unknown  $f$  that can be obtained from noisy evaluations of the  $f$  at fixed inputs (see Eq. (39)). Later we show that  $\Gamma_T$  can be bounded by a sum of the information gains for all  $m$  GPs. It is worth noting that Equation 17 may be a loose notion of model complexity because it assumes we can independently choose every  $\mathbf{z}_i$ , but for many graphs, there could be overlap in the  $\mathbf{z}_i, \mathbf{z}_j$  for  $i \neq j$  (two nodes could have a shared parent).

### A.2.2 ANALYSIS IN TERMS OF GENERAL MODEL COMPLEXITY $\Gamma_T$

In this section, we will prove a theorem similar to Theorem 1 but in terms of the model complexity we define in Eq. (17). Note that this version of the theorem does not require that  $\boldsymbol{\mu}$  and  $\boldsymbol{\sigma}$  come from a GP model with independent outputs, but any model such that Assumptions 1–3 are satisfied. In later sections, when using a GP model we bound  $\Gamma_T$  in terms of the maximum information gain of the  $m$  GPs used to get Theorem 1.

We will use the function  $\boldsymbol{\Sigma}_{i,t}(\cdot)$  to represent a matrix of all zeros except the values of the diagonal, given by  $\boldsymbol{\sigma}_{i,t}(\cdot)$ . As a result,  $\boldsymbol{\sigma}_{i,t}(\cdot) = \text{diag}(\boldsymbol{\Sigma}_{i,t}(\cdot))$

**Theorem 2.** *Consider the optimization problem in Eq. (4) with SCM satisfying Assumptions 1–3 where  $\mathcal{G}$  is known but  $\mathbf{F}$  is unknown. Then, for all  $T \geq 1$ , with probability at least  $1 - \delta$ , the regret of Algorithm 1 is bounded by*

$$R_T \leq \mathcal{O}\left(L_f^N L_\sigma^N \beta_T^N K^N \sqrt{Tm\Gamma_T}\right).$$

We first sketch the proof steps. In Lemma 1 we show that with high probability, there exists some set of functions  $\boldsymbol{\eta}$  that allows the reparameterized plausible SCM model in Eq. (11) to match the true SCM. Recall the mechanism of the ground-truth SCM in Eq. (2)

$$\mathbf{x}_{i,t} = \mathbf{f}_i(\mathbf{z}_{i,t}, \mathbf{a}_{i,t}) + \boldsymbol{\omega}_{i,t}, \quad \forall i \in \{0, \dots, m\}, \quad (2)$$

and the mechanism of the optimistic SCM model using the reparameterization of Eq. (11)

$$\tilde{\mathbf{x}}_{i,t} = \mathbf{f}_i(\tilde{\mathbf{z}}_{i,t}, \tilde{\mathbf{a}}_{i,t}) + \tilde{\boldsymbol{\omega}}_{i,t} \quad (18)$$

$$= \boldsymbol{\mu}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}, \tilde{\mathbf{a}}_{i,t}) + \beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}, \tilde{\mathbf{a}}_{i,t}) \boldsymbol{\eta}_i(\tilde{\mathbf{z}}_{i,t}, \tilde{\mathbf{a}}_{i,t}) + \tilde{\boldsymbol{\omega}}_{i,t}, \quad \forall i \in \{0, \dots, m\}. \quad (19)$$

In Lemma 2, Lemma 3 and Corollary 1 we bound the instantaneous regret (regret at some specific timepoint  $t$ ) by bounding the difference in SCM output, for the same action input, assuming the true SCM vs the optimistic reparameterized SCM. Then in Lemma 5 we use the intermediate result of Lemma 4 to show that our bound on instantaneous regret implies a bound on cumulative regret.

In Lemma 1, for convenience, we will drop the explicit dependence of all quantities on  $t$ .

**Lemma 1.** *Assume some fixed set of actions  $\mathbf{a}$  is chosen at any timepoint  $t$ . Under Assumption 3, for any  $\mathbf{x}$  generated by the true SCM Eq. (2), with probability at least  $1 - \delta$  there exists a set of functions  $\boldsymbol{\eta} = \{\eta_i\}_{i=0}^m$ , where  $\eta_i: \mathcal{Z}_i \times \mathcal{A}_i \rightarrow [-1, 1]^d$ , such that  $\mathbf{x} = \tilde{\mathbf{x}}$  if  $\forall i \ \boldsymbol{\omega}_i = \tilde{\boldsymbol{\omega}}_i$ .*

*Proof.* Since  $\boldsymbol{\omega} = \tilde{\boldsymbol{\omega}}$  we only need to prove that there exists some  $\boldsymbol{\eta}$  such that  $\forall i \ \mathbf{f}_i(\mathbf{z}_i, \mathbf{a}_i) = \boldsymbol{\mu}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) + \beta \boldsymbol{\Sigma}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) \boldsymbol{\eta}_i$ .

By Assumption 3, with probability  $1 - \delta$ , for all  $i = 0, \dots, m$ , we have an elementwise bound  $|\mathbf{f}_i(\mathbf{z}_i, \mathbf{a}_i) - \boldsymbol{\mu}_i(\mathbf{z}_i, \mathbf{a}_i)| \leq \beta_i \boldsymbol{\sigma}_i(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})$ . Thus for each  $\mathbf{z}_{i,t}, \mathbf{a}_{i,t}$  there exists a vector  $\boldsymbol{\eta}_i$  with values in  $[-1, 1]^d$  such that  $\mathbf{f}_i(\mathbf{z}_i, \mathbf{a}_i) = \boldsymbol{\mu}_i(\mathbf{z}_i, \mathbf{a}_i) + \beta \boldsymbol{\Sigma}_i(\mathbf{z}_i, \mathbf{a}_i) \boldsymbol{\eta}_i$ . Note that this is not quite what we need because the RHS contains  $\mathbf{z}_i$  and not  $\tilde{\mathbf{z}}_i$ . We will now use an inductive argument on  $i$  that constructs each  $\boldsymbol{\eta}_i$  sequentially from  $i = 0$  to  $m$ .

Base case: we must prove that for  $i = 0$  we have  $\mathbf{f}_0(\mathbf{z}_0, \mathbf{a}_0) = \boldsymbol{\mu}_0(\tilde{\mathbf{z}}_0, \tilde{\mathbf{a}}_0) + \beta_t \boldsymbol{\Sigma}_0(\tilde{\mathbf{z}}_0, \tilde{\mathbf{a}}_0) \boldsymbol{\eta}_0$ . We know  $\mathbf{z}_0 = \tilde{\mathbf{z}}_0$  (both are the empty vector) and  $\mathbf{a} = \tilde{\mathbf{a}}$  by the assumption of a fixed action. Then there exists some vector  $\boldsymbol{\eta}_0$  such that  $\mathbf{f}_0(\mathbf{z}_0, \mathbf{a}_0) = \boldsymbol{\mu}_0(\mathbf{z}_0, \mathbf{a}_0) + \beta \boldsymbol{\Sigma}_0(\mathbf{z}_0, \mathbf{a}_0) \boldsymbol{\eta}_0 = \boldsymbol{\mu}_0(\tilde{\mathbf{z}}_0, \tilde{\mathbf{a}}_0) + \beta \boldsymbol{\Sigma}_0(\tilde{\mathbf{z}}_0, \tilde{\mathbf{a}}_0) \boldsymbol{\eta}_0$ . Let  $\boldsymbol{\eta}_0(\cdot)$  be the function that outputs the vector  $\boldsymbol{\eta}_0$  given input  $\mathbf{z}_0, \mathbf{a}_0$  and the base case is proven.

Now assume the inductive hypothesis:  $\forall j < i$  we have  $\mathbf{f}_j(\mathbf{z}_j, \mathbf{a}_j) = \boldsymbol{\mu}_j(\tilde{\mathbf{z}}_j, \tilde{\mathbf{a}}_j) + \beta \boldsymbol{\Sigma}_j(\tilde{\mathbf{z}}_j, \tilde{\mathbf{a}}_j) \boldsymbol{\eta}_j$ . We want to show that this implies  $\mathbf{f}_i(\mathbf{z}_i, \mathbf{a}_i) = \boldsymbol{\mu}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) + \beta \boldsymbol{\Sigma}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) \boldsymbol{\eta}_i$ . We know  $\mathbf{a}_i = \tilde{\mathbf{a}}_i$  by the assumption of a fixed action.  $\mathbf{z}_i = \tilde{\mathbf{z}}_i$  because  $\tilde{\mathbf{z}}_i = [\tilde{\mathbf{x}}_{pa_i[1]}, \dots, \tilde{\mathbf{x}}_{pa_i[|pa_i|]}]^T$  and we selected each  $\boldsymbol{\eta}_j$  such that  $\tilde{\mathbf{x}}_j = \mathbf{x}_j$ . Then there exists some vector  $\boldsymbol{\eta}_i$  such that  $\mathbf{f}_i(\mathbf{z}_i, \mathbf{a}_i) = \boldsymbol{\mu}_i(\mathbf{z}_i, \mathbf{a}_i) + \beta \boldsymbol{\Sigma}_i(\mathbf{z}_i, \mathbf{a}_i) \boldsymbol{\eta}_i = \boldsymbol{\mu}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) + \beta \boldsymbol{\Sigma}_i(\tilde{\mathbf{z}}_i, \tilde{\mathbf{a}}_i) \boldsymbol{\eta}_i$ . Let  $\boldsymbol{\eta}_i(\cdot)$  output the vector  $\boldsymbol{\eta}_i$  given input  $\mathbf{z}_i, \mathbf{a}_i$  and the inductive step is proven.  $\square$

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

$$r_t = \mathbb{E}[y|\mathbf{F}, \mathbf{a}^*] - \mathbb{E}[y|\mathbf{F}, \mathbf{a}_{:,t}] \leq \mathbb{E}[y|\tilde{\mathbf{F}}_t, \mathbf{a}_{:,t}] - \mathbb{E}[y|\mathbf{F}, \mathbf{a}_{:,t}]. \quad (20)$$

*Proof.* The result follows directly from,

$$\mathbb{E}[y|\mathbf{F}, \mathbf{a}^*] \leq \mathbb{E}[y|\tilde{\mathbf{F}}_t, \mathbf{a}_{:,t}]. \quad (21)$$

This is true by the definition of  $\tilde{\mathbf{F}}_t, \mathbf{a}_{:,t}$  as the argmax of Eq. (12) and that with probability at least  $(1 - \delta)$  we have  $\mathbf{F} \in \mathcal{M}_T$ .  $\square$

We now show how the observations under the true and optimistic dynamics differ for a fixed noise sequence  $\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega}$  and the fixed action  $\mathbf{a}_{i,t}$  at any time  $t$ .

**Lemma 3.** *Under Assumptions 1–3, let  $\bar{L}_{f,t} = 1 + L_f + 2\beta_t L_\sigma$ . Then, for all iterations  $t > 0$ , any functions  $\boldsymbol{\eta}_i: \mathbb{R}^{p_i} \times \mathbb{R}^{q_i} \rightarrow [-1, 1]^{d_i}$  and any sequence of  $\boldsymbol{\omega}_i$  with  $\tilde{\boldsymbol{\omega}}_i = \boldsymbol{\omega}_i$  (for all  $i$ ), we have*

$$\|\mathbf{x}_{m,t} - \tilde{\mathbf{x}}_{m,t}\| \leq 2\beta_t K^N \bar{L}_{f,t}^N \sum_{i=0}^m \|\boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\| \quad (22)$$

*Proof.* We prove by induction on  $i$ .

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

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

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

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

$$\|\mathbf{x}_{i,t} - \tilde{\mathbf{x}}_{i,t}\| \stackrel{\textcircled{1}}{=} \|\mathbf{f}_i(\mathbf{z}_{i,t}) + \boldsymbol{\omega}_{i,t} - \boldsymbol{\mu}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) - \beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) \boldsymbol{\eta}_i(\tilde{\mathbf{z}}_{i,t}) - \tilde{\boldsymbol{\omega}}_{i,t}\|$$

$$\begin{aligned}
& \stackrel{(2)}{=} \|\mathbf{f}_i(\mathbf{z}_{i,t}) - \boldsymbol{\mu}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) - \beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) \boldsymbol{\eta}_i(\tilde{\mathbf{z}}_{i,t}) + \mathbf{f}_i(\tilde{\mathbf{z}}_{i,t}) - \mathbf{f}_i(\tilde{\mathbf{z}}_{i,t})\| \\
& \stackrel{(3)}{=} \|\mathbf{f}_i(\tilde{\mathbf{z}}_{i,t}) - \boldsymbol{\mu}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) - \beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) \boldsymbol{\eta}_i(\tilde{\mathbf{z}}_{i,t}) + \mathbf{f}_i(\mathbf{z}_{i,t}) - \mathbf{f}_i(\tilde{\mathbf{z}}_{i,t})\| \\
& \stackrel{(4)}{\leq} \|\mathbf{f}_i(\tilde{\mathbf{z}}_{i,t}) - \boldsymbol{\mu}_{i,t-1}(\tilde{\mathbf{z}}_{i,t})\| + \|\beta_t \boldsymbol{\Sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) \boldsymbol{\eta}_i(\tilde{\mathbf{z}}_{i,t})\| + \|\mathbf{f}_i(\mathbf{z}_{i,t}) - \mathbf{f}_i(\tilde{\mathbf{z}}_{i,t})\| \\
& \stackrel{(5)}{\leq} \beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t})\| + \beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t})\| + L_f \|\mathbf{z}_{i,t} - \tilde{\mathbf{z}}_{i,t}\| \\
& \stackrel{(6)}{=} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t})\| + L_f \|\mathbf{z}_{i,t} - \tilde{\mathbf{z}}_{i,t}\| \\
& \stackrel{(7)}{=} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\tilde{\mathbf{z}}_{i,t}) + \boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t}) - \boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t})\| + L_f \|\mathbf{z}_{i,t} - \tilde{\mathbf{z}}_{i,t}\| \\
& \stackrel{(8)}{\leq} 2\beta_t (\|\boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t})\| + L_\sigma \|\mathbf{z}_{i,t} - \tilde{\mathbf{z}}_{i,t}\|) + L_f \|\mathbf{z}_{i,t} - \tilde{\mathbf{z}}_{i,t}\| \\
& \stackrel{(9)}{\leq} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t})\| + (1 + L_f + 2\beta_t L_\sigma) \|\mathbf{z}_{i,t} - \tilde{\mathbf{z}}_{i,t}\| \\
& \stackrel{(10)}{\leq} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t})\| + (1 + L_f + 2\beta_t L_\sigma) \sum_{j \in pa_i} \|\mathbf{x}_{i,t} - \tilde{\mathbf{x}}_{i,t}\| \\
& \stackrel{(11)}{\leq} 2\beta_t \|\boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t})\| \\
& + (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)^{N_j}}_{=: L_f} \sum_{h=0}^j \|\boldsymbol{\sigma}_{h,t-1}(\mathbf{z}_{h,t})\| \\
& \stackrel{(12)}{\leq} 2\beta_t K^{N_i} \bar{L}_{f,t}^{N_i} \sum_{j=0}^i \|\boldsymbol{\sigma}_{j,t-1}(\mathbf{z}_{j,t})\| \tag{23}
\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 we can 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  $N_j = N_i - 1$ .  $\square$

Now we will relate this bound on the observations to a bound on  $y_t$  when selecting actions according to MCBO in both the optimistic and true dynamics.

**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*

$$\mathbb{E} \left[ y_t | \tilde{\mathbf{F}}_t, \mathbf{a}_{:,t} \right] - \mathbb{E} [y_t | \mathbf{F}, \mathbf{a}_{:,t}] \leq 2\beta_t K^N \bar{L}_{f,t}^N \mathbb{E}_{\omega=\tilde{\omega}} \left[ \sum_{i=0}^m \|\boldsymbol{\sigma}_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\| \right] \tag{24}$$

*Proof.* This follows from Lemma 3.  $Y$  is just the final observation so

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

$\square$



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

$$r_t^2 \leq L_{Y,t}^2 K^{2N} m \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\|_2^2 \right] \quad (25)$$

*Proof.*

$$r_t \leq \mathbb{E} [y_t | \tilde{\mathbf{F}}_t, \mathbf{a}_t] - \mathbb{E} [y_t | \mathbf{F}, \mathbf{a}_t] \quad (26)$$

$$\leq 2\beta_t K^N \bar{L}_{f,t}^N \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\| \right] \quad (27)$$

$$\leq L_{Y,t} K^N \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\| \right] \quad (28)$$

$$r_t^2 \leq L_{Y,t}^2 K^{2N} \left( \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\| \right] \right)^2 \quad (29)$$

$$\leq L_{Y,t}^2 K^{2N} \mathbb{E} \left[ \left( \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\| \right)^2 \right] \quad (30)$$

$$\leq L_{Y,t}^2 K^{2N} m \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\|_2^2 \right] \quad (31)$$

The last two lines are Jensen's inequality.  $\square$

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

$$R_T^2 \leq T L_{Y,T}^2 m \sum_{t=1}^T \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\|_2^2 \right] \quad (32)$$

*Proof.*

$$R_T^2 = \left( \sum_{t=1}^T r_t \right)^2 \stackrel{\textcircled{1}}{\leq} T \sum_{t=1}^T r_t^2 \stackrel{\textcircled{2}}{\leq} T L_{Y,T}^2 m \sum_{t=1}^T \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\|_2^2 \right], \quad (33)$$

$\square$

where  $\textcircled{1}$  is due to Jensen's inequality and  $\textcircled{2}$  follows Lemma 4. Similar to the equivalent lemma in 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  $\Gamma_T$ .

**Lemma 6.** Under Assumptions 1–3, with probability at least  $(1 - \delta)$  it holds for all  $t \geq 0$  that

$$R_T^2 \leq T L_{Y,T}^2 K^{2N} m \Gamma_T \quad (34)$$

*Proof.* Substituting in Eq. (17) we have

$$\sum_{t=1}^T \mathbb{E} \left[ \sum_{i=0}^m \|\sigma_{i,t}(\mathbf{z}_{i,t}, \mathbf{a}_{i,t})\|_2^2 \right] \leq \Gamma_T \quad (35)$$

and the result follows.  $\square$

Taking square roots and substituting in for  $L_{Y,T}$  in terms of  $\beta_T$ ,  $L_f$  and  $L_\sigma$  in Eq. (34) concludes the proof for Theorem 2.

## A.2.3 PROOF OF THEOREM 1

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

**Theorem 1.** *Consider the optimization problem in Eq. (4) with SCM satisfying Assumptions 1–3 where  $\mathcal{G}$  is known but  $\mathbf{F}$  is unknown. Then for all  $T \geq 1$ , with probability at least  $1 - \delta$ , the cumulative regret of Algorithm 1 is bounded by*

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

*Proof. Step 1.* Some preliminaries relating mutual information  $I_T(\mathbf{x}_{i,1:T}, f_{i,1:T})$  and maximum information gain  $\gamma_T$

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

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

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

$$\begin{aligned} \boldsymbol{\mu}_{t-1}(\mathbf{z}_t, \mathbf{a}_t) &= [\mu_{t-1}(\mathbf{z}_t, \mathbf{a}_t, 1), \dots, \mu_{t-1}(\mathbf{z}_t, \mathbf{a}_t, d)]^\top, \\ \boldsymbol{\Sigma}_{t-1}(\mathbf{z}_t, \mathbf{a}_t) &= \text{diag}(\boldsymbol{\Sigma}_{t-1}(\mathbf{z}_t, \mathbf{a}_t)) = [\sigma_{t-1}(\mathbf{z}_t, \mathbf{a}_t, 1), \dots, \sigma_{t-1}(\mathbf{z}_t, \mathbf{a}_t, d)]^\top. \end{aligned}$$

Our setup assumes that the components  $\{\mathbf{x}_{t,l}\}_{l=0}^d$  are independent of each other given  $\mathbf{z}_t, \mathbf{a}_t$ . Therefore from Srinivas et al. (2010) we know that the mutual information for the  $i$ th GP model is:

$$I(\mathbf{x}_{1:T}, \mathbf{f}_{1:T}) = H(\mathbf{x}_{1:T}) - H(\mathbf{x}_{1:T} | \mathbf{f}_{1:T}) = \frac{1}{2} \sum_{t=1}^T \sum_{l=1}^d \ln \left( 1 + \frac{\sigma_{t-1}^2(\mathbf{z}_t, \mathbf{a}_t, l)}{\rho^2} \right) \quad (36)$$

because per component:

$$I(\mathbf{x}_{i,1:T,l}, \mathbf{f}_{i,1:T,l}) = \frac{1}{2} \sum_{t=1}^T \ln \left( 1 + \frac{\sigma_{t-1}^2(\mathbf{z}_t, \mathbf{a}_t, l)}{\rho^2} \right). \quad (37)$$

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

$$\gamma_{i,T,l} := \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} I(\mathbf{x}_{i,1:T,l}, \mathbf{f}_{i,1:T,l}) = \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} \frac{1}{2} \sum_{t=1}^T \ln \left( 1 + \frac{\sigma_{i,t-1}^2(\mathbf{z}_t, \mathbf{a}_t, l)}{\rho^2} \right) \quad (38)$$

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

$$\gamma_{i,T} := \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} I(\mathbf{x}_{i,1:T}, \mathbf{f}_{i,1:T}) = \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} \frac{1}{2} \sum_{t=1}^T \sum_{l=1}^d \ln \left( 1 + \frac{\sigma_{i,t-1}^2(\mathbf{z}_t, \mathbf{a}_t, l)}{\rho^2} \right) \quad (39)$$

$$\leq \sum_{l=1}^d \max_{A_i \subset \{\mathcal{Z}_i \times \mathcal{A}_i\}^T} \frac{1}{2} \sum_{t=1}^T \ln \left( 1 + \frac{\sigma_{i,t-1}^2(\mathbf{z}_t, \mathbf{a}_t, l)}{\rho^2} \right) = \sum_{l=1}^d \gamma_{i,T,l} \quad (40)$$

**Note:** Bounds for  $\gamma_{i,T,l}$  and  $\gamma_{i,T}$ . Upper bounds on  $\gamma_{i,T,l}$  are provided in Srinivas et al. (2010) for widely used kernels and scale sublinearly in  $T$ . We use  $p_i = d |pa(i)|$  to represent the size

of the  $z$ -input to a GP. Recall that  $q$  is the length of each action vector i.e.  $\mathcal{A}_i \subset \mathbb{R}^q$ . For the linear kernel  $\gamma_{i,T,l} = \mathcal{O}((p_i + q) \log T)$ , and for the squared exponential kernel  $\gamma_{i,T,l} = \mathcal{O}((p_i + q)(\log T)^{p_i+q+1})$ . We can use Eq. (39) to give bounds on  $\gamma_{i,T}$  too e.g. if all GPs use independent linear kernels for each output component  $\gamma_{i,T} = \mathcal{O}(d_i(p_i + q) \log T)$ , and if all GPs use independent squared exponential kernels for each output component  $\gamma_{i,T} = \mathcal{O}(d(p_i + q)(\log T)^{p_i+q+1})$ .

**Step 2. A bound for model complexity  $\Gamma_T$ .**

Here we bound the model complexity Eq. (17)

$$\Gamma_T \leq \sum_{i=0}^m \frac{1}{\ln(1 + \rho_i^{-2})} \gamma_{i,T}. \quad (41)$$

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

$$\begin{aligned} \Gamma_T &= \max_{\{\mathcal{Z} \times \mathcal{A}\}^T} \sum_{t=1}^T \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i)\|_2^2 \\ &\stackrel{\textcircled{1}}{\leq} \max_A \sum_{t=1}^T \sum_{i=0}^m \|\sigma_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i)\|_2^2 \stackrel{\textcircled{2}}{\leq} \sum_{i=0}^m \max_A \sum_{t=1}^T \|\sigma_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i)\|_2^2 \\ &\stackrel{\textcircled{3}}{\leq} \sum_{i=0}^m \max_{A_i} \sum_{t=1}^T \|\sigma_{i,t-1}(\mathbf{z}_i, \mathbf{a}_i)\|_2^2 \stackrel{\textcircled{4}}{\leq} \sum_{i=0}^m \max_{A_i} \sum_{t=1}^T \sum_{l=1}^{d_i} \|\sigma_{i,(t-1)}(\mathbf{z}_i, \mathbf{a}_i, l)\|_2^2 \\ &\stackrel{\textcircled{5}}{\leq} \sum_{i=0}^m \frac{2}{\ln(1 + \rho_i^{-2})} \underbrace{\max_{A_i} \frac{1}{2} \sum_{t=1}^T \sum_{l=1}^{d_i} \ln \left( 1 + \frac{\sigma_{i,(t-1)}(\mathbf{z}_i, \mathbf{a}_i, l)}{\rho_i^2} \right)}_{\text{maximum information gain Eq. (39)}} \\ &\stackrel{\textcircled{6}}{=} \sum_{i=0}^m \frac{2}{\ln(1 + \rho_i^{-2})} \gamma_{i,T} \stackrel{\textcircled{7}}{=} \mathcal{O}(m\gamma_T), \end{aligned}$$

where  $\textcircled{1}$  bounds  $\mathcal{A}$  and  $\mathcal{Z}$  with a box.  $\textcircled{2}$  is from the max over a sum.  $\textcircled{3}$  is due to the assumption of  $\mathbf{x}_{i,t}$  being independent of  $A_j$ ,  $j \neq i$ , conditioned on  $A_i$ .  $\textcircled{4}$  is due to Jensen's inequality.  $\textcircled{5}$  is due to the fact that for any  $s^2 \in [0, \rho^{-2}]$  we can bound  $s^2 \leq \frac{\rho^{-2}}{\ln(1+\rho^{-2})} \ln(1+s^2)$  Srinivas et al. (2010). This also holds for function  $s^2(\cdot) := \rho_i^{-2} \sigma_{i,(t-1),l}^2(\cdot)$  since  $\rho_i^{-2} \sigma_{i,(t-1),d}^2(\cdot) \leq \rho_i^{-2} k(\cdot, \cdot) \leq \rho_i^{-2}$  because  $k_i(\cdot, \cdot) < 1 \forall i$  (bounded kernel assumption).  $\textcircled{6}$  is due to Eqs. (36) and (39). Finally, in  $\textcircled{7}$  we define  $\gamma_T := \max_i \gamma_{i,T}$  being the maximum value of the maximum information gains over graph nodes.

Plugging this upper bound on  $\Gamma_T$  into Theorem 2 completes the proof.

**Note:** Sublinearity w.r.t.  $T$  of maximum information gain  $\gamma_T$ .

Upper bounds on  $\gamma_T$  will often scale sublinearly in  $T$ . This follows from  $\gamma_{i,T}$  scaling sublinearly in  $T$  for many popularly used kernels (see previous note). In particular, a linear kernel leads to  $\gamma_T = \mathcal{O}(d(Kd + q) \log T)$  and a squared exponential kernel leads to  $\gamma_T = \mathcal{O}(d(Kd + q)(\log T)^{Kd+q+1})$  (assuming output components are independent) since  $\max_i p_i = Kd$ . □

#### A.2.4 DEPENDENCE OF $\beta_T$ ON $T$ FOR PARTICULAR KERNELS

Note that for Assumption 3 to hold under our RKHS assumptions,  $\beta_T$  might depend on  $T$ . For a single observed variable corresponding to node  $i$  at time  $t$ , for Assumption 3 to hold we must have  $\beta_T = \tilde{\mathcal{O}}(\mathcal{B}_i + \frac{\rho_i}{d} \sqrt{\gamma_{i,t}})$  (Chowdhury & Gopalan (2019) equation 9). For Assumption 3 to hold at

time  $t$  for all  $i$  we can apply a union bound and see that it is sufficient for  $\beta_T = \tilde{\mathcal{O}}\left(\mathcal{B} + \frac{\rho}{d}\sqrt{\gamma_t}\right)$  where  $\mathcal{B} = \max_i \mathcal{B}_i$  and  $\rho = \max_i \rho_i$ .

In the regret bound of Theorem 1,  $\beta_T$  appears raised to the power  $N$ . For  $\gamma_T$  corresponding to the linear kernel and squared exponential kernel, this will still lead to sublinear regret regardless of  $N$  because  $\gamma_T$  will be only logarithmic in  $T$ . However, for a Matérn kernel, where the best known bound  $\gamma_T = \mathcal{O}(p(pT)^c \log(pT))$  with  $0 < c < 1$ , the cumulative regret bound will not be sublinear if  $N$  and  $c$  are sufficiently large. A similar phenomena with the Matérn kernel appears in the guarantees of Curi et al. (2020) which use GP models in model-based reinforcement learning.

### A.3 MAXIMIZING THE ACQUISITION FUNCTION

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  $\eta_i$  as a neural network, we can parameterize it as a constant. This is because with no noise, the inputs to  $\eta_i$  ( $z_i$  and  $a_i$ ) given  $a$  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 (2021b) which is also an out-of-the-box optimizer in the BoTorch package Balandat et al. (2020).

For noisy settings where each  $\eta_i : \mathcal{A}_i \times \mathcal{Z}_i \rightarrow \mathbb{R}$  is a neural network, we use our own optimizer. For each initialization of  $\eta$  parameters, we perform stochastic gradient descent to optimize both the  $\eta_i$  parameters and  $a$ . We can do this because Eq. (12) is differentiable with respect to both  $a$  and the parameters of each  $\eta_i$ . After running stochastic gradient descent on many random initializations we will have many solution candidates. We select the candidate 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) for model-based reinforcement learning, could also be adapted to optimize our acquisition function.

When parameterizing each  $\eta_i$  with a neural network, we always use a two layer feed-forward network with a ReLU non-linearity, To map the output into  $[-1, 1]$  we put the output of the network through an element-wise Sigmoid.

In all noisy environments we estimate the expectation in the acquisition function (Eq. (12)) using a Monte Carlo estimate with 32 repeats for each gradient step. For noisy Dropwave we use 128 repeats because the environment is particularly noisy compared to other noisy environments.

### A.4 EXPERIMENTAL SETUP

Here we give additional notes to explain the details of our experimental setup.

The cross-validation for selecting  $\beta$  is performed across  $\beta = \{0.05, 0.5, 5\}$ . For a given performance metric and task, we'll use the example of average reward on Dropwave, we select the  $\beta$  that on average performs best across all other tasks for the same performance metric, called  $\beta^*$ . We then report the results in terms of average reward for running the algorithm (GP-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 (2021b). 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 the agent selects no intervention (observational data) at a given timestep, 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 to make the environment more stable.

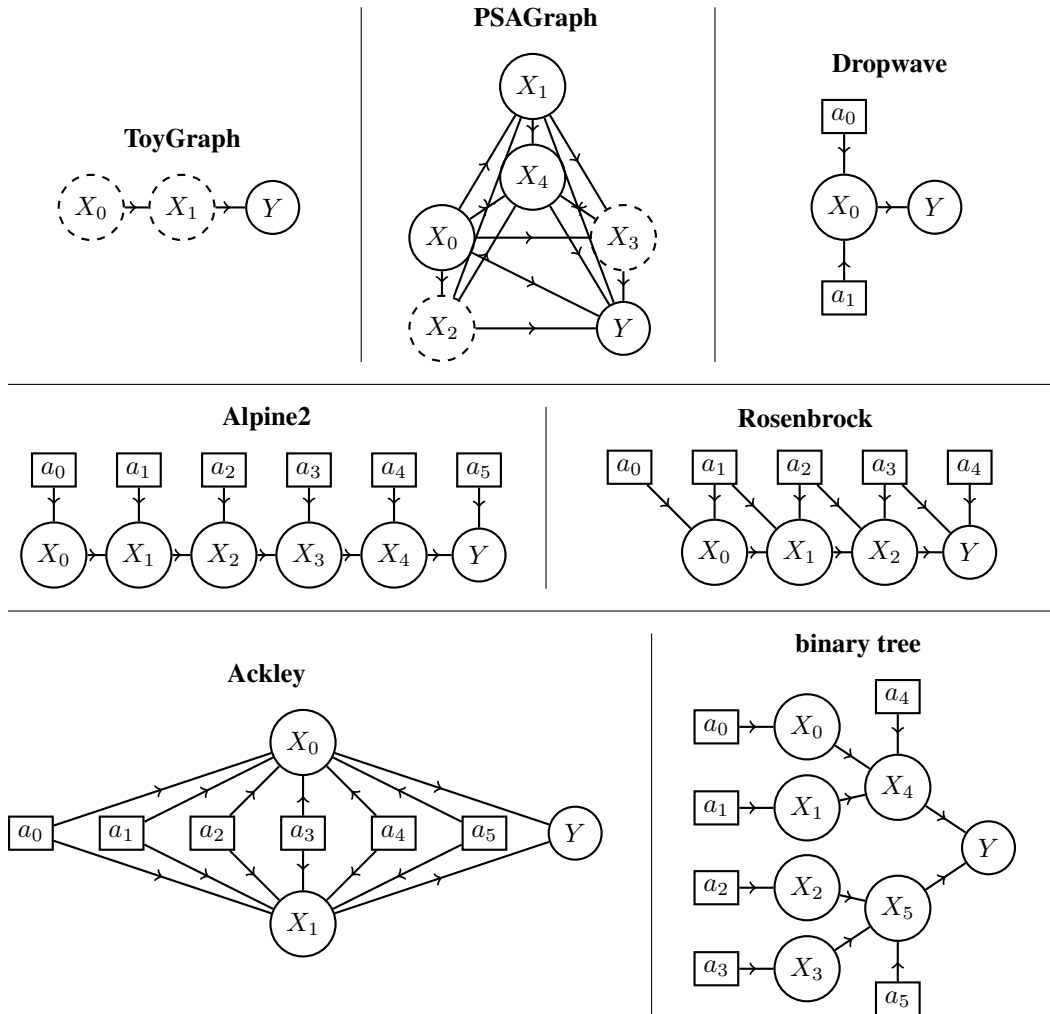


Figure 3: The DAGs corresponding to each task in the experiments, except binary tree which is used for illustrative purposes. Circles are observation or reward variables. Squares are actions in the function network setting. In hard intervention CBO, nodes with a dashed border can be intervened upon. All nodes represent a scalar random variable.

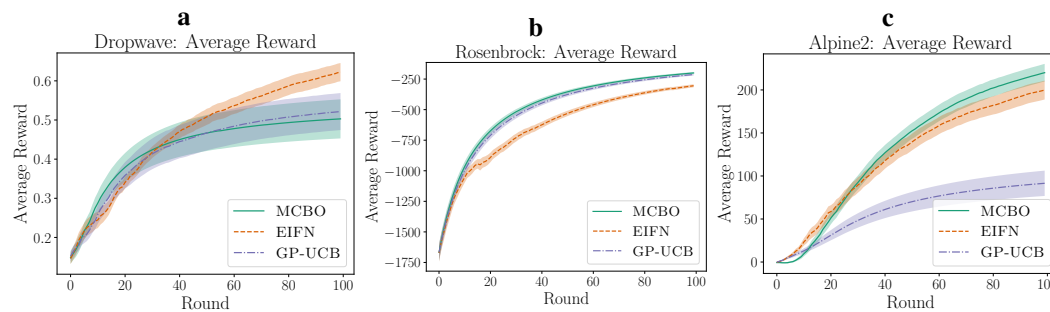


Figure 4: All average reward plots not used in the main paper. All settings are noiseless.

When applying MCBO to function networks we allow  $a_m$  to be nonzero, to match the function networks setting.

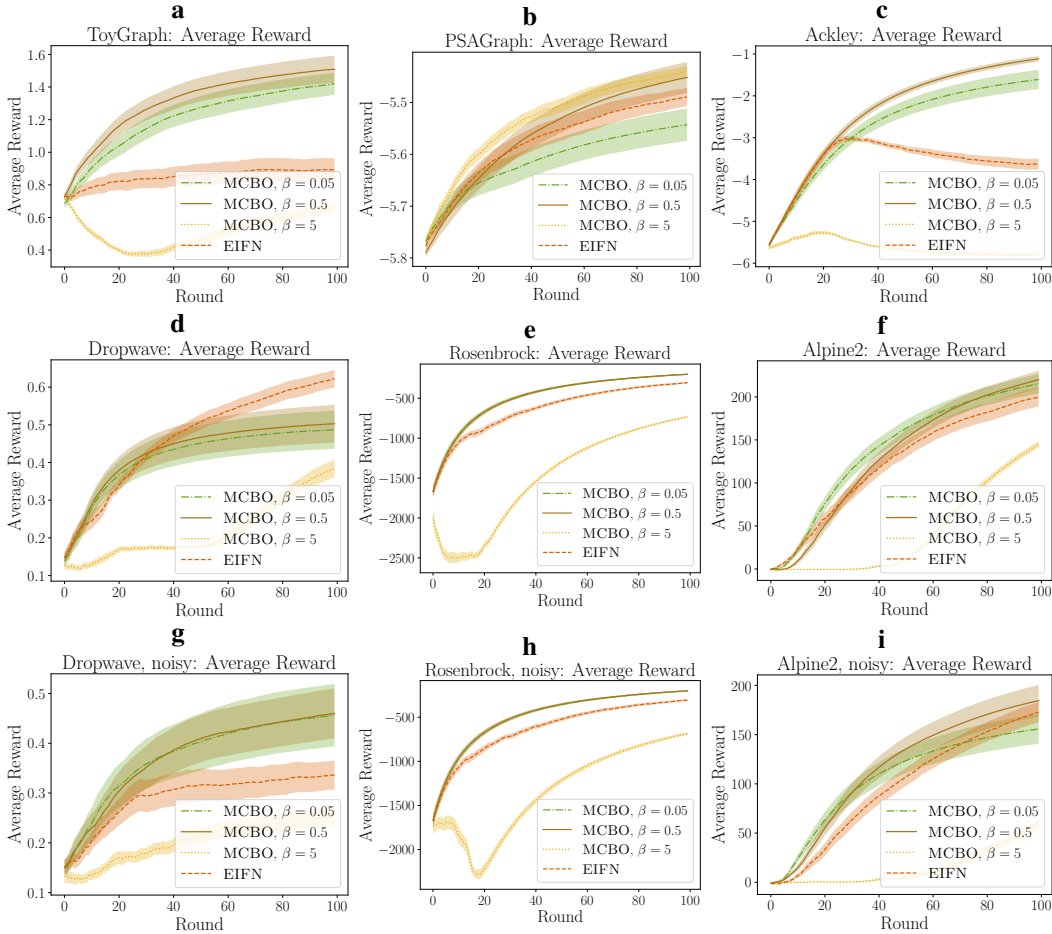


Figure 5: An ablation across  $\beta$  where we plot average reward for all tasks.

### A.5 FURTHER EXPERIMENTAL ANALYSIS

#### Average expected reward

In Figure 4 we show the deterministic function networks plots not included in the main paper. In Figure 5 we show the average reward for all  $\beta$  considered in the cross-validation.

**Best expected reward** Here we also report the best expected reward as a function of the number of rounds  $T$ . The best expected reward at time  $T$  is defined by

$$\max_{\mathbf{a} \in \{\mathbf{a}_t\}_{t=0}^T} \mathbb{E}_\omega[Y | \mathbf{a}].$$

This is similar to but not directly inversely related to simple regret. Simple regret algorithms require an inference procedure to select a final action after  $T$  rounds of exploration. This procedure could be computationally expensive for the CBO setting. For example, a reasonable choice would be to report a final action based upon maximizing a lower confidence bound of the objective in Eq. (4), however a closed form expression for this does not exist for CBO. Best expected reward instead assumes access to an oracle that computes the highest expected reward of any action the algorithm has previously chosen. Our plotting of the best expected reward is consistent with previous work on CBO (Aglietti et al., 2020b).

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

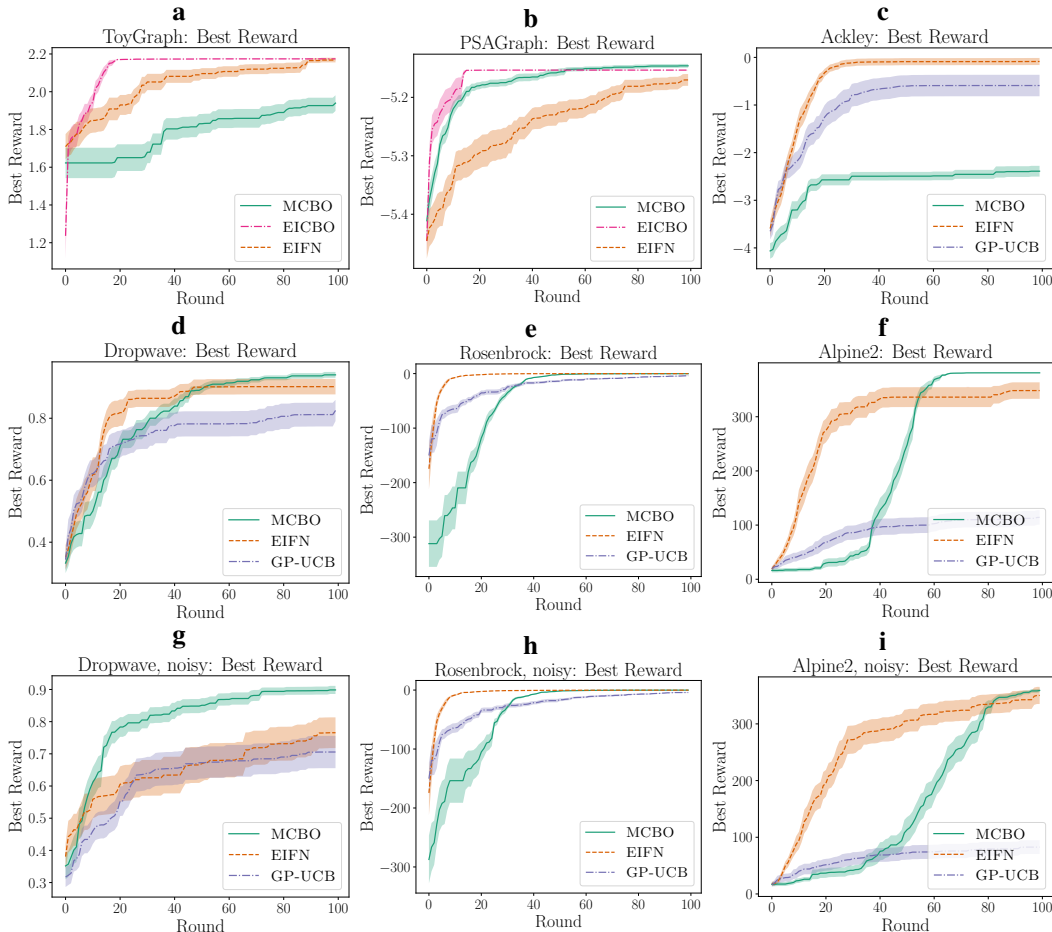


Figure 6: All best reward plots.

the best  $\beta$  among the 3 tried, we find that MCBO is very strong on all tasks (with the exception of best reward on ToyGraph, where it underperforms the baselines). The optimal  $\beta$  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 cumulative regret guarantee in Theorem 1.

The plots of average and best scores for other values of  $\beta$  (Figs. 5 and 7) suggest that  $\beta$  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 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 GP-UCB and EICBO. MCBO generally has a much longer runtime than EIFN in noisy settings where the  $\eta_i$  are parameterized by neural networks. Generally in BO, because the unknown function is often assumed expensive to evaluate, we are less concerned about the time required to optimize the acquisition function and more concerned with notions of statistical efficiency such as regret. Our implementation for noisy settings could likely be sped-up with more parallelism to make it more 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 Fig. 2 (a,b) we found that EICBO had a non-monotonic average reward, which could likely be explained by the use of an expected improvement acquisition function. We decided to clarify that the non-monotonic average reward was because of the algorithm and not because of differences in our setup compared to Aglietti et al. (2020b). Besides us evaluat-

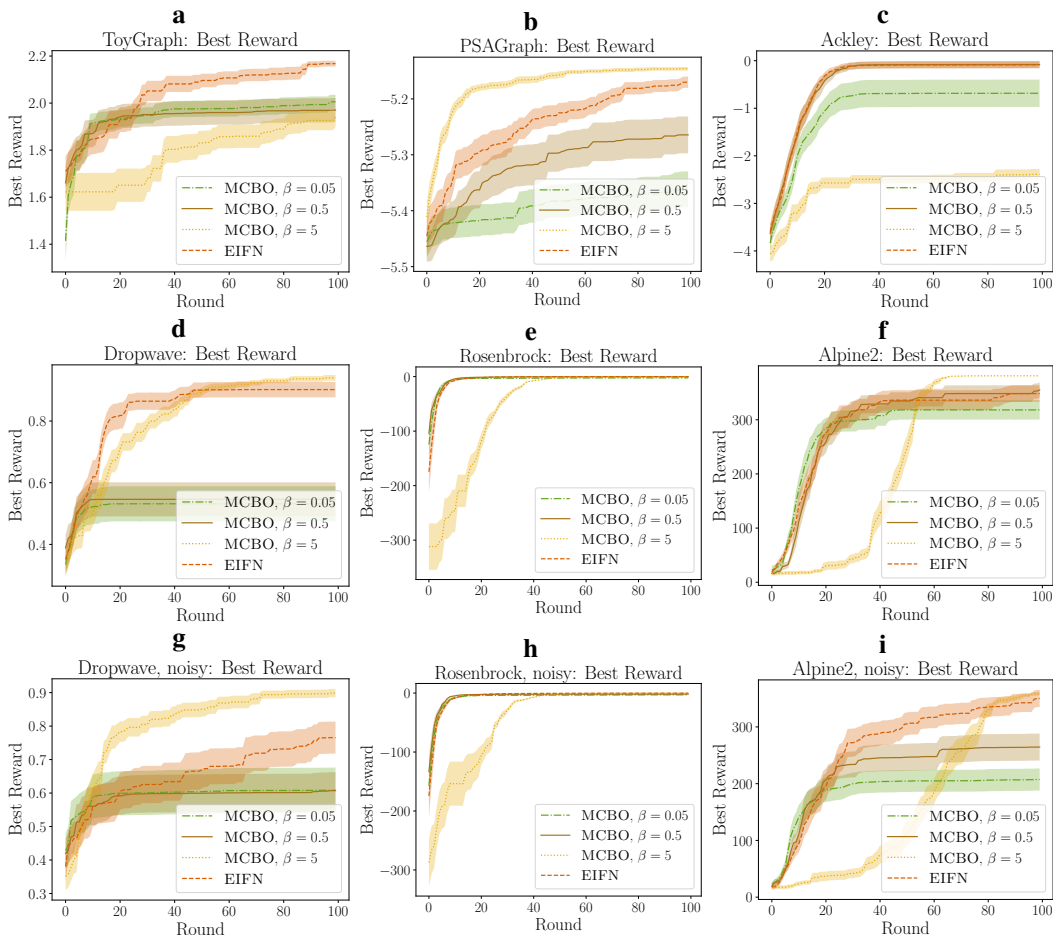


Figure 7: An ablation across  $\beta$  where we plot best reward for all tasks.

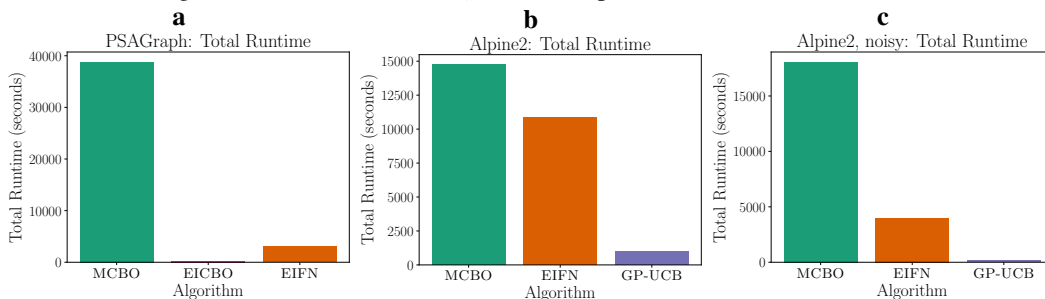


Figure 8: Runtimes of experiments for a) CBO setting, b) noiseless function network setting, and c) noisy function network setting.

ing in terms of average reward instead of best reward, compared to Aglietti et al. (2020b) we also used smaller initial sample sizes (random samples the agent obtains before  $t = 0$ ) of observational data. This was to make the setting more challenging, since we found that with extra observational data ToyGraph and PSAGraph could be solved with very few samples. When reproducing the experiments of Fig. 2 (a, b) with additional starting observational data, we found the same qualitative results (non-monotonicity) for the average reward case. Note that this result is not inconsistent at all with the experimental results of Aglietti et al. (2020b), since they only evaluate in terms of best reward.