
Information-Theoretic Bayesian Optimization for Bilevel Optimization Problems

Takuya Kanayama¹

Yuki Ito¹

Tomoyuki Tamura²

Masayuki Karasuyama¹

¹Dept. of Computer Science, Nagoya Institute of Technology, Nagoya, Aichi, Japan

²Dept. of Physical Science and Engineering, Nagoya Institute of Technology, Nagoya, Aichi, Japan

Abstract

A bilevel optimization problem consists of two optimization problems nested as an upper- and a lower-level problem, in which the optimality of the lower-level problem defines a constraint for the upper-level problem. This paper considers Bayesian optimization (BO) for the case that both the upper- and lower-levels involve expensive black-box functions. Because of its nested structure, bilevel optimization has a complex problem definition, by which bilevel BO has not been widely studied compared with other standard extensions of BO such as multi-objective or constraint problems. We propose an information-theoretic approach that considers the information gain of both the upper- and lower-optimal solutions and values. This enables us to define a unified criterion that measures the benefit for both level problems, simultaneously. Further, we also show a practical lower bound based approach to evaluating the information gain. We empirically demonstrate the effectiveness of our proposed method through several benchmark datasets.

1 INTRODUCTION

The bilevel optimization is a standard formulation for a decision making problem that has a hierarchical structure. It consists of two optimization problems nested as an upper- and a lower-level problem, in which the optimality of the lower-level problem defines a constraint for the upper-level problem. Bilevel optimization techniques is applicable to hierarchical decision makings in a variety of contexts such as inverse optimal control [Suryan et al., 2016], chemical reaction optimization [Abbassi et al., 2021], and shape optimization [Herskovits et al., 2000].

We particularly focus on the case that both level problems

are defined by expensive black-box functions, while most of existing studies assume that the lower level problem is not expensive to observe. For example, consider the case that the upper- and lower-objective functions are defined through simulators of a subject of interest, which can occur in a variety of scientific, engineering, and industrial fields. If these simulators consist of expensive computations (such as quantum-mechanical calculations), both level problems are expensive to observe. An example is that the simulator-based optimization of a physical property of inorganic crystals (computational materials design) under the stability constraint (energy minimization) can be seen as an instance of this class of problems.

Most of existing BO studies for bilevel optimization applies BO only to the upper-level problem [e.g., Kieffer et al., 2017, Dogan and Prestwich, 2023] as pointed out by [Chew et al., 2025]. Typically, for a selected query to the upper-level problem, repeated queries to the lower-level problem is required, and/or further, the gradient of the lower-level problem is often assumed [e.g., Fu et al., 2024]. A few studies [Islam et al., 2018, Wang et al., 2021] consider BO in both levels, but repeated queries on the lower-level is still required. These approaches are not fully suitable when both level problems are expensive black-boxes in which the gradient is not available. It is well-known that hyper-parameter optimization (HPO) of a machine learning model can be seen as a bilevel problem. However, the lower objective is usually inexpensive white-box function, and therefore, HPO is different from our focus, while there exist specific studies for HPO bilevel problems [e.g., Lorraine et al., 2020]. Further, some recent studies assume that the objectives are defined as an average (expectation) of some base functions and the stochastic samples of it are available, which can occur in problems such as HPO, federated learning, and adversarial attacks [e.g., Aghasi and Ghadimi, 2025, Nazari et al., 2025, Qiu et al., 2023, Jiao et al., 2025], while our framework do not employ this assumption (further, in these papers, objectives are not necessarily expensive).

On the other hand, recently a few methods have been stud-

ied for the both level expensive setting. Ekmekcioglu et al. [2024] combine the Thompson sampling on the upper-level query and a knowledge gradient-based extension of multi-task BO on the lower-level, but a rationale behind the combination of two different criteria has not been revealed. Further, Chew et al. [2025] propose the well-known GP upper confidence bound (UCB) based approach to bilevel BO, called BILBO. Although BILBO has a theoretical regret guarantee, in general, the performance of GP-UCB based methods depend on the selection of the balancing parameter of the exploitation and exploration, because the theoretically recommended value often does not provide the best performance [Srinivas et al., 2010].

We propose an information-theoretic approach that considers the simultaneous information gain for both the upper- and lower-optimal solutions and values, which we call bilevel information gain. This enables us to define a unified criterion that measures the benefit for both level problems simultaneously, which is not necessarily common in the case of bilevel methods as already mentioned. Although the effectiveness of information-theoretic BO has been shown in several different contexts [e.g., Hennig and Schuler, 2012, Hernández-Lobato et al., 2014, Hoffman and Ghahramani, 2015, Wang and Jegelka, 2017, Hernández-Lobato et al., 2015, Hernandez-Lobato et al., 2016, Suzuki et al., 2020, Takeno et al., 2022a,b, Hvarfner et al., 2022, Tu et al., 2022], it has not been combined with bilevel optimization, to our knowledge. We first define bilevel information gain by extending the idea of the joint entropy search [Hvarfner et al., 2022, Tu et al., 2022]. Unfortunately, the original definition of bilevel information gain is computationally intractable, and we show that a natural extension of the truncation based approximation, which has been widely employed in information-theoretic BO [e.g., Wang and Jegelka, 2017], can be derived. By combining the truncation based approximation and a variational lower bound [Takeno et al., 2022b], we obtain our criterion called Bilevel optimization via Lower-bound based Joint Entropy Search (BLJES).

Further, while we mainly consider ‘coupled setting’ in which upper- and lower-level observations are obtained simultaneously, ‘decoupled setting’, in which a separate observation for each level is available, is also discussed. For example, in the case that the objective function values are outputs of some simulators (e.g., physical simulation), if a common simulator provides both level observations simultaneously, coupled setting is suitable, while if the upper- and the lower-level observations are from different simulators, decoupled setting can be more appropriate. We further propose an extension for the case that each level problem has inequality constraints (i.e., each level problem is a constraint problem).

Our contributions are summarized as follows.

- We show an information-theoretic formulation of

bilevel BO, which has never been explored, to our knowledge. Bilevel information gain is defined to measure the benefit for both level problems.

- We derive a lower bound based approximation of bilevel information gain. We extend the standard truncation based approach in the single-level information-theoretic BO to the bilevel problem.
- We further propose extensions for decoupled setting and constraint problems. We show that our framework can handle these settings by a natural extension of bilevel information gain.

We demonstrate effectiveness of BLJES through functions generated from Gaussian processes and several benchmark problems.

2 PRELIMINARIES

Bilevel optimization. Let $f : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ and $g : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ denote the upper- and the lower-level objective functions, respectively, both of which are assumed to be costly black-box functions. The upper- and the lower-level variables are denoted by $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$, respectively, where $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\Theta \subset \mathbb{R}^{d_\theta}$. The bilevel optimization problem is formulated as:

$$\begin{aligned} \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \\ \text{s.t. } \boldsymbol{\theta}^*(\mathbf{x}) = \arg \max_{\boldsymbol{\theta} \in \Theta} g(\mathbf{x}, \boldsymbol{\theta}), \end{aligned} \quad (1)$$

where $\boldsymbol{\theta}^*(\mathbf{x})$ represents the optimal solution of the lower-level problem for a given upper-level variable $\mathbf{x}$. For simplicity, we assume the lower-level optimum $\boldsymbol{\theta}^*(\mathbf{x})$ is uniquely determined for each $\mathbf{x}$ [Zhang et al., 2024]. The bilevel optimal solution is denoted by $(\mathbf{x}^*, \boldsymbol{\theta}^*)$, while the lower-level optimum corresponding to a given $\mathbf{x}$ is written as $(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))$, noting that $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\mathbf{x}^*)$. The upper- and the lower-level optimal values are denoted by $f^* := f(\mathbf{x}^*, \boldsymbol{\theta}^*)$ and $g^* := g(\mathbf{x}^*, \boldsymbol{\theta}^*)$, respectively. Figure 1 shows an illustration. Observations of the objective functions contain additive Gaussian noise $y_{(\mathbf{x}, \boldsymbol{\theta})}^f := f(\mathbf{x}, \boldsymbol{\theta}) + \epsilon^f$, $\epsilon^f \sim \mathcal{N}(0, \{\sigma_{\text{noise}}^f\}^2)$, and $y_{(\mathbf{x}, \boldsymbol{\theta})}^g := g(\mathbf{x}, \boldsymbol{\theta}) + \epsilon^g$, $\epsilon^g \sim \mathcal{N}(0, \{\sigma_{\text{noise}}^g\}^2)$, where σ_{noise}^f and σ_{noise}^g are the noise standard deviations, respectively. Let $\mathcal{D}_t := \{(\mathbf{x}_i, \boldsymbol{\theta}_i, y_i^f, y_i^g)\}_{i=1}^n$ be the dataset that we have at the t -th iteration of BO, where $y_i^f := y_{(\mathbf{x}_i, \boldsymbol{\theta}_i)}^f$ and $y_i^g := y_{(\mathbf{x}_i, \boldsymbol{\theta}_i)}^g$. The number of observed points n is $t + n_0$, where n_0 is the number of the initial observations.

Gaussian process. The upper- and the lower-level objective functions are each modeled by independent Gaussian processes (GPs) with kernel functions $k^f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'))$ and $k^g((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'))$, respectively. Given the dataset $\mathcal{D}_t$, the predictive distribution of an objective function

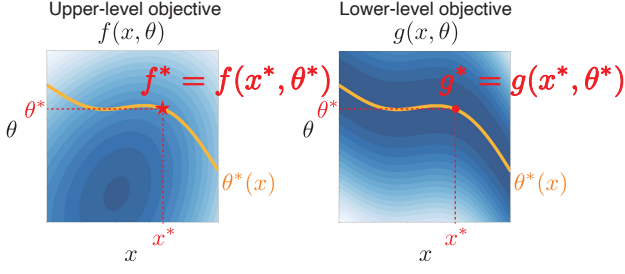


Figure 1: Example of Bilevel optimization ($d_{\mathcal{X}} = 1, d_{\Theta} = 1$) and its optimal solution. For each upper-level variable x , the feasible solution is defined by $\theta^*(x) = \arg \max_{\theta} g(x, \theta)$, shown as the orange line. The optimal solution (x^*, θ^*) is the maximizer of f on the orange line.

$h \in \{f, g\}$ at a point (x, θ) is expressed as:

$$\begin{aligned} h(x, \theta) | \mathcal{D}_t &\sim \mathcal{N}(\mu_t^h(x, \theta), \{\sigma_t^h(x, \theta)\}^2), \\ \mu_t^h(x, \theta) &= \mathbf{k}^h \top (\mathbf{K}^h + \{\sigma_{\text{noise}}^h\}^2 \mathbf{I})^{-1} \mathbf{y}^h, \\ \{\sigma_t^h(x, \theta)\}^2 &= \mathbf{k}^h((x, \theta), (x, \theta)) \\ &\quad - \mathbf{k}^h \top (\mathbf{K}^h + \{\sigma_{\text{noise}}^h\}^2 \mathbf{I})^{-1} \mathbf{k}^h, \end{aligned} \quad (2)$$

where $\mathbf{k}^h = (k^h((x, \theta), (x_1, \theta_1)), \dots, k^h((x, \theta), (x_t, \theta_n)))^\top$, $\mathbf{y}^h = (y_1^h, \dots, y_n^h)^\top$, and $\mathbf{K}^h \in \mathbb{R}^{n \times n}$ is the kernel matrix with an entry $k^h((x_i, \theta_i), (x_j, \theta_j))$ at a position (i, j) . Here, $\mathbf{I} \in \mathbb{R}^{n \times n}$ denotes the identity matrix.

Bayesian optimization. We consider BO for the bilevel optimization problem (1). Bayesian optimization is a method for efficiently optimizing black-box functions with a limited number of samples. At step t , GPs are fitted to the dataset $\mathcal{D}_t$, and the next query point is determined as $\arg \max_{x, \theta} \alpha_t(x, \theta)$, where $\alpha_t(x, \theta)$ denotes the acquisition function. After sampling the query point, the newly obtained data are added to the dataset, and the GPs are refitted.

3 BILEVEL OPTIMIZATION VIA LOWER-BOUND BASED JOINT ENTROPY SEARCH

We consider bilevel BO based on the information gain for the optimal solutions and values $(x^*, \theta^*, f^*, \text{ and } g^*)$ achieved by next observations $y_{(x, \theta)}^f$ and $y_{(x, \theta)}^g$, which we call bilevel information gain. Note that we regard the optimal $(x^*, \theta^*, f^*, g^*)$ as random variables defined by the predictive distributions of the objective functions $f(x, \theta)$ and $g(x, \theta)$. Our approach combines the concept of entropy search [Hennig and Schuler, 2012], in particular, joint entropy search [Hvarfner et al., 2022, Tu et al., 2022], and a variational lower bound based approximation of mutual information (MI). We refer to our proposed method as *Bilevel*

optimization via Lower-bound based Joint Entropy Search (BLJES).

3.1 LOWER BOUND OF MUTUAL INFORMATION

Bilevel information gain is represented as the MI between the candidate observations $(y_{(x, \theta)}^f, y_{(x, \theta)}^g)$ and the set of the optimal solutions and their upper- and lower-objective values $o^* := \{f^*, g^*, x^*, \theta^*\}$:

$$\text{MI}(y_{(x, \theta)}^f, y_{(x, \theta)}^g; o^* | \mathcal{D}_t),$$

for which an illustration is shown in Fig. 2. This criterion naturally allows simultaneous consideration of both the upper- and the lower-objectives. Since the direct evaluation of this MI is difficult, we employ a lower bound based approximation. Let $\Omega := \{y_{(x, \theta)}^f, y_{(x, \theta)}^g, f^*, g^*, x^*, \theta^*\}$. Our lower bound of the MI is derived by a technique that is often used in the context of the variational approximation [e.g., Poole et al., 2019]:

$$\begin{aligned} &\text{MI}(y_{(x, \theta)}^f, y_{(x, \theta)}^g; o^* | \mathcal{D}_t) \\ &= \mathbb{E}_{\Omega} \left[\log \frac{p(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t)}{p(y_{(x, \theta)}^f, y_{(x, \theta)}^g | \mathcal{D}_t)} \right] \\ &= \mathbb{E}_{o^*} \left[\mathbb{E}_{y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t} \left[\log \frac{q(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t)}{p(y_{(x, \theta)}^f, y_{(x, \theta)}^g | \mathcal{D}_t)} \right] \right. \\ &\quad \left. + \text{KL} \left(p(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t) \parallel q(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t) \right) \right] \\ &\geq \mathbb{E}_{\Omega} \left[\log \frac{q(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t)}{p(y_{(x, \theta)}^f, y_{(x, \theta)}^g | \mathcal{D}_t)} \right] =: \text{LB}(x, \theta), \end{aligned} \quad (3)$$

where KL is Kullback-Leibler (KL) divergence and $q(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t)$ is a variational distribution (q can be any density function as far as the KL divergence can be defined). The inequality of the last line can be taken because the KL divergence is non-negative (the equality holds when $p(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t) = q(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t)$). Similar lower bounds of the MI have been used in information-theoretic multi-objective and constraint BO [Ishikura and Karasuyama, 2025, Takeno et al., 2022b].

The variational distribution q is an approximation of $p(y_{(x, \theta)}^f, y_{(x, \theta)}^g | o^*, \mathcal{D}_t)$ for which an exact analytical representation is difficult to know. The difficulty is in the conditioning by the optimal solutions and values, for which the most widely accepted approach in information-theoretic BO is to use truncated distributions [e.g., Wang and Jegelka, 2017, Suzuki et al., 2020, Tu et al., 2022]. For example, in the case of well-known max-value entropy search (MES), proposed by [Wang and Jegelka, 2017] for the standard single-level problem $\max_x f(x)$, the predictive distribution conditioning on the max-value $f_{\text{unc}}^* := \arg \max_x f(x)$ is approximated by the truncated normal distribution, i.e.,

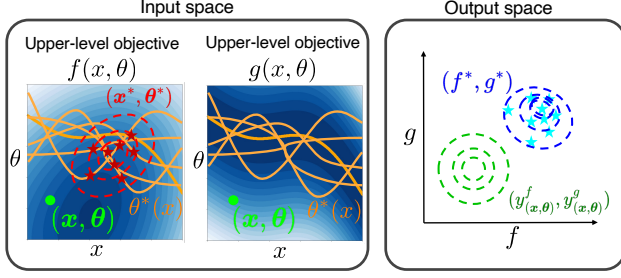


Figure 2: Schematic illustration of random variables in $\text{MI}(y^f_{(x,\theta)}, y^g_{(x,\theta)}; o^* | \mathcal{D}_t)$. The left panel represents input space of f and g (heatmap is the true functions), and the right panel is the output space. The predictions at current (x, θ) are $y^f_{(x,\theta)}$ and $y^g_{(x,\theta)}$, depicted as the green distribution in the output space. At the optimal solution (x^*, θ^*) , which is also a random variable defined through the GP posteriors, the objective function values are $f^* = f(x^*, \theta^*)$ and $g^* = g(x^*, \theta^*)$. The optimal values f^* and g^* are represented as the blue distribution in the output space. The optimal solution (x^*, θ^*) , represented as the red distribution in the input space, should exist on $\theta^*(x)$, i.e., the optimal point of the lower-level problem. The red and blue stars are ‘samples’ of the optimal solutions (x^*, θ^*) and values (f^*, g^*) , respectively.

$p(f(x) | f^*_{\text{unc}}) \approx p(f(x) | f(x) \leq f^*_{\text{unc}})$. When f^*_{unc} is given, $f(x') \leq f^*_{\text{unc}}$ should hold for any x' (and there should exist at least one x' such that $f(x') = f^*_{\text{unc}}$), while MES simplifies this condition so that $f(x) \leq f^*_{\text{unc}}$ holds only for the current x . Similar simplifications have been employed by most of information-theoretic BO algorithms and shown superior performance.

We extend the truncation based approach to our bilevel problem as follows.

$$q(y^f_{(x,\theta)}, y^g_{(x,\theta)} | o^*, \mathcal{D}_t) := p(y^f_{(x,\theta)}, y^g_{(x,\theta)} | f(x, \theta^*(x)) \leq f^*, g(x^*, \theta) \leq g^*, \mathcal{D}_t^+), \quad (4)$$

where $\mathcal{D}_t^+ = \mathcal{D}_t \cup \{(x^*, \theta^*, f^*, g^*)\}$ is the dataset augmented by the optimal point $(x^*, \theta^*, f^*, g^*)$. The right hand side has the three conditions, each of which can be interpreted as follows.

- When f^* is given, $f(x', \theta^*(x')) \leq f^*$ should hold for $\forall x'$. However, this condition is computationally intractable as mentioned for the case of MES. Based on a similar idea to MES, the condition $f(x, \theta^*(x)) \leq f^*$ is only imposed on the current x .
- When g^* is given, $g(x^*, \theta') \leq g^*$ should hold for $\forall \theta'$. We replace it with $g(x^*, \theta) \leq g^*$ in which the inequality is only imposed on the current θ .
- In the right hand side of (4), $\mathcal{D}_t$ is replaced with $\mathcal{D}_t^+$. This condition imposes that the GPs satisfy

$f(x^*, \theta^*) = f^*$ and $g(x^*, \theta^*) = g^*$ by adding a noiseless observation $(x^*, \theta^*, f^*, g^*)$ into the training data.

By substituting (4) into (3) and using the conditional independence of $y^f_{(x,\theta)}$ and $y^g_{(x,\theta)}$ in the right hand side of (4) (posteriors of independent GPs are also independent), we see

$$\begin{aligned} \text{LB}(x, \theta) &= \mathbb{E}_{\Omega} \left[\log \frac{p(y^f_{(x,\theta)}, y^g_{(x,\theta)} | f(x, \theta^*(x)) \leq f^*, g(x^*, \theta) \leq g^*, \mathcal{D}_t^+)}{p(y^f_{(x,\theta)}, y^g_{(x,\theta)} | \mathcal{D}_t)} \right] \\ &= \mathbb{E}_{\Omega} \left[\log \frac{p(y^f_{(x,\theta)} | f(x, \theta^*(x)) \leq f^*, \mathcal{D}_t^+)}{p(y^f_{(x,\theta)} | \mathcal{D}_t)} \right. \\ &\quad \left. + \log \frac{p(y^g_{(x,\theta)} | g(x^*, \theta) \leq g^*, \mathcal{D}_t^+)}{p(y^g_{(x,\theta)} | \mathcal{D}_t)} \right]. \quad (5) \end{aligned}$$

The inside of the expectation (5) can be analytically derived. For both the first and second terms, the denominators are the predictive distribution of the GPs, whose density can be obtained from (2). Next, we consider the numerator of the first term of (5). Although the truncation is imposed on $f(x, \theta^*(x))$, the distribution is for $y^f_{(x,\theta)}$ that has different input point (x, θ) from the truncated point, unlike the case of MES. The following theorem shows that the analytical representation can still be derived even with this difference:

Theorem 3.1. Let (m_1^f, s_1^f) , (m_2^f, s_2^f) , and (m_3^f, s_3^f) be the mean and standard deviation of $p(f(x, \theta^*(x)) | y^f_{(x,\theta)}, \mathcal{D}_t^+)$, $p(f(x, \theta^*(x)) | \mathcal{D}_t^+)$, and $p(y^f_{(x,\theta)} | \mathcal{D}_t^+)$, respectively. Then,

$$p(y^f_{(x,\theta)} | f(x, \theta^*(x)) \leq f^*, \mathcal{D}_t^+) = \begin{cases} \frac{\Phi\left(\frac{f^* - m_1^f}{s_1^f}\right) \phi\left(\frac{y^f_{(x,\theta)} - m_3^f}{s_3^f}\right)}{\Phi\left(\frac{f^* - m_2^f}{s_2^f}\right) s_3^f} & \text{if } x \neq x^*, \\ \phi\left(\frac{y^f_{(x,\theta)} - m_3^f}{s_3^f}\right) / s_3^f & \text{otherwise,} \end{cases} \quad (6)$$

where ϕ and Φ are the probability density function (PDF) and cumulative density function (CDF) of the standard normal distribution, respectively.

The proof is in Appendix A.1. Note that all of $\{(m_i^f, s_i^f)\}_{i=1}^3$ can be analytically calculated from the GP posterior of f easily, for which details are also show in Appendix A.1. For the numerator in the second term of (5) can be reduced to the similar analytical form, which is shown in Appendix A.2. As a result, we obtain an analytical form of the inside of the expectation (5).

3.2 COMPUTATIONS

The expectation of (5) is approximated by the Monte-Carlo method by sampling $\Omega := \{y^f_{(x,\theta)}, y^g_{(x,\theta)}, f^*, g^*, x^*, \theta^*\}$.

The sample of all the elements of Ω can be obtained through the sample of the objective functions f and g . We use random Fourier feature (RFF) [Rahimi and Recht, 2008], by which the GP posterior can be approximated by the Bayesian linear model. RFF based approximation has been repeatedly used by information-theoretic BO methods [e.g., Shah and Ghahramani, 2015]. For a D -dimensional RFF vector $\phi(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^D$, the linear model $\mathbf{w}^{h\top} \phi(\mathbf{x}, \boldsymbol{\theta})$ can be constructed, where $\mathbf{w}^h \in \mathbb{R}^D$ is a parameter vector and $h \in \{f, g\}$ represents one of objective functions. By sampling $\mathbf{w}^h$ from the posterior, approximate sample paths of f and g are obtained, which are denoted as $\tilde{f}$ and $\tilde{g}$, respectively. Then, the sample of $(f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*)$ is obtained through $\max_{\mathbf{x}} \tilde{f}(\mathbf{x}, \tilde{\boldsymbol{\theta}}^*(\mathbf{x}))$ s.t. $\tilde{\boldsymbol{\theta}}^*(\mathbf{x}) = \arg \max_{\boldsymbol{\theta}} \tilde{g}(\mathbf{x}, \boldsymbol{\theta})$, which can be seen as a white-box bilevel optimization problem. Since both $\tilde{f}$ and $\tilde{g}$ are represented by the Bayesian linear model, they are differentiable. Then, the gradient $\partial \tilde{f}(\mathbf{x}, \tilde{\boldsymbol{\theta}}^*(\mathbf{x})) / \partial \mathbf{x}$ can be obtained through the implicit function theorem (see Appendix B for detail), by which standard gradient based optimization methods can be applied. The samples of $y_{(\mathbf{x}, \boldsymbol{\theta})}^f$ and $y_{(\mathbf{x}, \boldsymbol{\theta})}^g$ can be obtained by adding the random noise from $\mathcal{N}(0, \{\sigma_{\text{noise}}^f\}^2)$ and $\mathcal{N}(0, \{\sigma_{\text{noise}}^g\}^2)$ to the sampled $f(\mathbf{x}, \boldsymbol{\theta})$ and $g(\mathbf{x}, \boldsymbol{\theta})$, respectively.

Let K be the number of samplings of Ω . In a variety of contexts of information-theoretic BO [e.g., Wang and Jegelka, 2017], the superior performance has been repeatedly shown with small K settings (e.g., 10). After obtaining K samples of Ω , the Monte-Carlo approximation of (5) can be calculated for any $(\mathbf{x}, \boldsymbol{\theta})$ (Note that these K samples are reused during the acquisition function optimization without regenerating for each candidate $(\mathbf{x}, \boldsymbol{\theta})$). Since (5) contains $\boldsymbol{\theta}^*(\mathbf{x})$, the maximization of the approximate (5) is also bilevel optimization, for which gradient-based methods can also be applied through the implicit function theorem (details are also in Appendix B). The outline of the algorithm of BLJES is illustrated in Algorithm 1.

4 EXTENSIONS

We here describe two extensions of BLJES, which are for decoupled setting and constraint problems.

4.1 DECOUPLED SETTING

We mainly consider the setting in which $y_{(\mathbf{x}, \boldsymbol{\theta})}^f$ and $y_{(\mathbf{x}, \boldsymbol{\theta})}^g$ are observed simultaneously, which we call ‘coupled’ setting. On the other hand, only one of $y_{(\mathbf{x}, \boldsymbol{\theta})}^f$ or $y_{(\mathbf{x}, \boldsymbol{\theta})}^g$ can be separately observed in some scenarios. In this paper, this setting is called ‘decoupled’ setting, inspired by the similar setting in multi-objective BO [Hernandez-Lobato et al., 2016].

A natural criterion for decoupled setting is information gain

Algorithm 1: BLJES

```

1 function BLJES( $\mathcal{D}_0, K, T$ )
2   for  $t = 1, \dots, T$  do
3     for  $k = 1, \dots, K$  do
4       Generate posterior sample paths of  $f$  and  $g$ 
5       Sample  $\Omega$  by solving white-box bilevel
6         problem defined by sampled  $f$  and  $g$ 
7       Calculate inside of expectation (5) by using
8         sampled  $\Omega$ 
9     end
10     $a \leftarrow K$ -sample MC approximation of lower
11      bound function (5)
12     $(\mathbf{x}_{t+1}, \boldsymbol{\theta}_{t+1}) \leftarrow \arg \max_{\mathbf{x}, \boldsymbol{\theta}} a(\mathbf{x}, \boldsymbol{\theta})$ 
13    Query  $(\mathbf{x}_{t+1}, \boldsymbol{\theta}_{t+1})$ , and observe  $(y_{t+1}^f, y_{t+1}^g)$ 
14     $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup (\mathbf{x}_{t+1}, \boldsymbol{\theta}_{t+1}, y_{t+1}^f, y_{t+1}^g)$ 
15  end

```

obtained by only one of $y_{(\mathbf{x}, \boldsymbol{\theta})}^f$ or $y_{(\mathbf{x}, \boldsymbol{\theta})}^g$, for which the lower bounds can be derived by the almost same way as (3):

$$\begin{aligned} & \text{MI}(y_{(\mathbf{x}, \boldsymbol{\theta})}^f ; f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^* | \mathcal{D}_t) \\ & \geq \mathbb{E}_{\Omega} \left[\log \frac{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f | f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*, \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f | \mathcal{D}_t)} \right], \end{aligned} \quad (7)$$

$$\begin{aligned} & \text{MI}(y_{(\mathbf{x}, \boldsymbol{\theta})}^g ; f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^* | \mathcal{D}_t) \\ & \geq \mathbb{E}_{\Omega} \left[\log \frac{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g | g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g | \mathcal{D}_t)} \right]. \end{aligned} \quad (8)$$

The derivation is in Appendix C. For both the inside of the expectation of (7) and (8), the analytical calculations shown in section 3.1 can be used. The expectation is approximated by Monte-Carlo sampling of Ω , which is also same as coupled setting. As a result, the decision making not only for selecting $(\mathbf{x}, \boldsymbol{\theta})$, but also selecting the upper- or the lower-observation (i.e., $y_{(\mathbf{x}, \boldsymbol{\theta})}^f$ or $y_{(\mathbf{x}, \boldsymbol{\theta})}^g$) can be performed.

4.2 INCORPORATING CONSTRAINT PROBLEMS

In a more general formulation of bilevel optimization, constraints are imposed on both of the upper- and the lower-level problems. When we have N and M inequality constraints for the upper- and the lower-level problems, respectively, the bilevel optimization problem is written as

$$\begin{aligned} & \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \\ & \text{s.t. } c_n^U(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \geq 0, n = 1, \dots, N \\ & \quad \boldsymbol{\theta}^*(\mathbf{x}) = \arg \max_{\boldsymbol{\theta} \in \Theta} \{g(\mathbf{x}, \boldsymbol{\theta}) \mid c_m^L(\mathbf{x}, \boldsymbol{\theta}) \geq 0, m = 1, \dots, M\}, \end{aligned}$$

where $c^U : \mathbb{R}^{d_{\mathcal{X}} + d_{\Theta}} \rightarrow \mathbb{R}$ and $c^L : \mathbb{R}^{d_{\mathcal{X}} + d_{\Theta}} \rightarrow \mathbb{R}$ are constraint functions. We assume that c^U and c^L are also

expensive black-box functions and modeled by the independent GPs.

For constraint BO, Takeno et al. [2022b] show an information-theoretic approach based on a lower bound with a truncated variational distribution. By combining the truncation shown by [Takeno et al., 2022b] and our bilevel information gain, we can extend BLJES to the bilevel constraint problem. In particular, the conditioning on the predictive distributions by the optimal points are required to extend. For example, if f^* is given, the inequality $f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*$ is imposed only when the constraints $c_n^U(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \geq 0, n = 1, \dots, N$ hold (if the constraints are not satisfied, $f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))$ is not truncated). Details are in Appendix D.

5 EXPERIMENTS

We evaluated the performance of BLJES by using sample path functions from the GP prior and several benchmark functions. For baselines, we employed Random selection and BILBO. The initial number of observations was set $n_0 = 5$ random points. The both level observations contain an additive noise whose mean is 0 and standard deviation is 10^{-3} . Each experiment was performed 10 times with different initial points. We used the Gaussian kernel for both the GPs of f and g , in which the prior mean, the kernel length-scale, the output scale, and the noise variance are optimized by the marginal likelihood at every iteration. In BLJES, the number of Mont-Carlo samples was set as $K = 30$. We here employed the pool setting (query candidates are finite grid points defined later) because BILBO is proposed for the finite domain setting.

For the metric at the t -th iteration, we used the following criterion, denoted as bilevel simple regret:

$$\min_{i \in [n_0+t]} \max_{h \in \{f, g\}} r_h(\mathbf{x}_i, \boldsymbol{\theta}_i), \quad (9)$$

where

$$r_f(\mathbf{x}_i, \boldsymbol{\theta}_i) = \frac{\max(0, f^* - f(\mathbf{x}_i, \boldsymbol{\theta}_i))}{(f^* - \min_{\mathbf{x}, \boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}))},$$

$$r_g(\mathbf{x}_i, \boldsymbol{\theta}_i) = \frac{(g(\mathbf{x}_i, \boldsymbol{\theta}^*(\mathbf{x}_i)) - g(\mathbf{x}_i, \boldsymbol{\theta}_i))}{(g(\mathbf{x}_i, \boldsymbol{\theta}^*(\mathbf{x}_i)) - \min_{\boldsymbol{\theta}} g(\mathbf{x}_i, \boldsymbol{\theta}))}.$$

Our metric (9) takes the larger value between $r_f(\mathbf{x}_i, \boldsymbol{\theta}_i)$, which represents the regret of the upper-level problem, and $r_g(\mathbf{x}_i, \boldsymbol{\theta}_i)$, which represents those of the lower-level problem. Since $f(\mathbf{x}_i, \boldsymbol{\theta}_i)$ can be larger than f^* , the ‘max’ operation is taken to guarantee $r_f(\mathbf{x}_i, \boldsymbol{\theta}_i) \geq 0$, while the numerator of $r_g(\mathbf{x}_i, \boldsymbol{\theta}_i)$ is non-negative without ‘max’ from the definition of $\boldsymbol{\theta}^*(\mathbf{x}_i)$. The denominators of $r_f(\mathbf{x}_i, \boldsymbol{\theta}_i)$ and $r_g(\mathbf{x}_i, \boldsymbol{\theta}_i)$ are for absorbing the scale difference of two objectives. In (9), we employed the best value obtained during the entire search procedure by taking the minimum with respect to observed points.

We first provide the results on coupled setting for GP sample path functions (section 5.1) and benchmark and real-world functions (section 5.2). Further, the results on decoupled setting (section 5.3) and different K settings (section 5.4) are also reported. Appendix presents other details of the settings (Appendix E.1) and other empirical evaluations such as a comparison with the standard EI (Appendix E.2), a larger noise setting (Appendix E.3), the continuous domain (Appendix E.6), constraint problems (Appendix E.7), higher dimensional settings (Appendix E.8), comparison with a simplified variant of MLJES (Appendix E.9), the effect of approximations (Appendix E.10), and computational time of acquisition function (Appendix E.11).

5.1 SAMPLE PATH FROM GP PRIOR

We first used the sample path from the GP prior as the true objective functions, i.e., $f \sim \mathcal{GP}(0, k)$ and $g \sim \mathcal{GP}(0, k)$, in which k is the Gaussian kernel $k((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}')) = \exp\{-(\|\mathbf{x} - \mathbf{x}'\|^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2)/(2\ell^2)\}$. For the length scale ℓ , we use different values $\ell_U \in \{0.25, 0.10, 0.50\}$ for f and $\ell_L \in \{0.25, 0.10, 0.50\}$ for g , respectively. The input space is $d_{\mathcal{X}} = d_{\Theta} = 1$ and $[0, 1]$ for each. The candidate points are a combination of 100 grid points in each dimension (100^2 points).

The results are shown in Fig. 3. Overall, BLJES shows superior performance for a variety of the length scale functions. Only for $(\ell_U, \ell_L) = (0.25, 0.50)$, BILBO decreased the regret to 0 faster at about 60 iterations, but BLJES also reached the small value (10^{-4}) around that iterations.

5.2 BENCHMARK AND REAL-WORLD DATA

We here used the five benchmark and three real-world problems. Two functions are created by combining benchmark functions of single-level optimization. In the first problem, denoted as BG, the upper objective is BraninHoo ($d_{\mathcal{X}} = 1$) and the lower objective Goldstein-price ($d_{\Theta} = 1$), which was used in [Chew et al., 2025]. In the second problem, denoted as SB, the upper objective is SixHumpCamel ($d_{\mathcal{X}} = 1$) and the lower objective BraninHoo ($d_{\Theta} = 1$), which was used in [Ekmekcioglu et al., 2024]. From the third to the fifth problems, denoted as SMD01, 02, and 03 ($d_{\mathcal{X}} = 2$ and $d_{\Theta} = 2$), are test problems specifically designed for bilevel optimization benchmark [Sinha et al., 2014]. The sixth problem, denoted as Energy, is a simulator based energy market problem ($d_{\mathcal{X}} = 2$ and $d_{\Theta} = 2$), and the seventh problem, denoted as Chemical, is about an optimization of simulated mass flow of Methyl Acetate ($d_{\mathcal{X}} = 1$ and $d_{\Theta} = 3$). Chemical has one constraint function for the upper-level problem. Energy and Chemical data are introduced by [Chew et al., 2025], in which these are regarded as a real-world dataset (see Chew et al. [2025] for the detailed definitions). The number of grid points in each

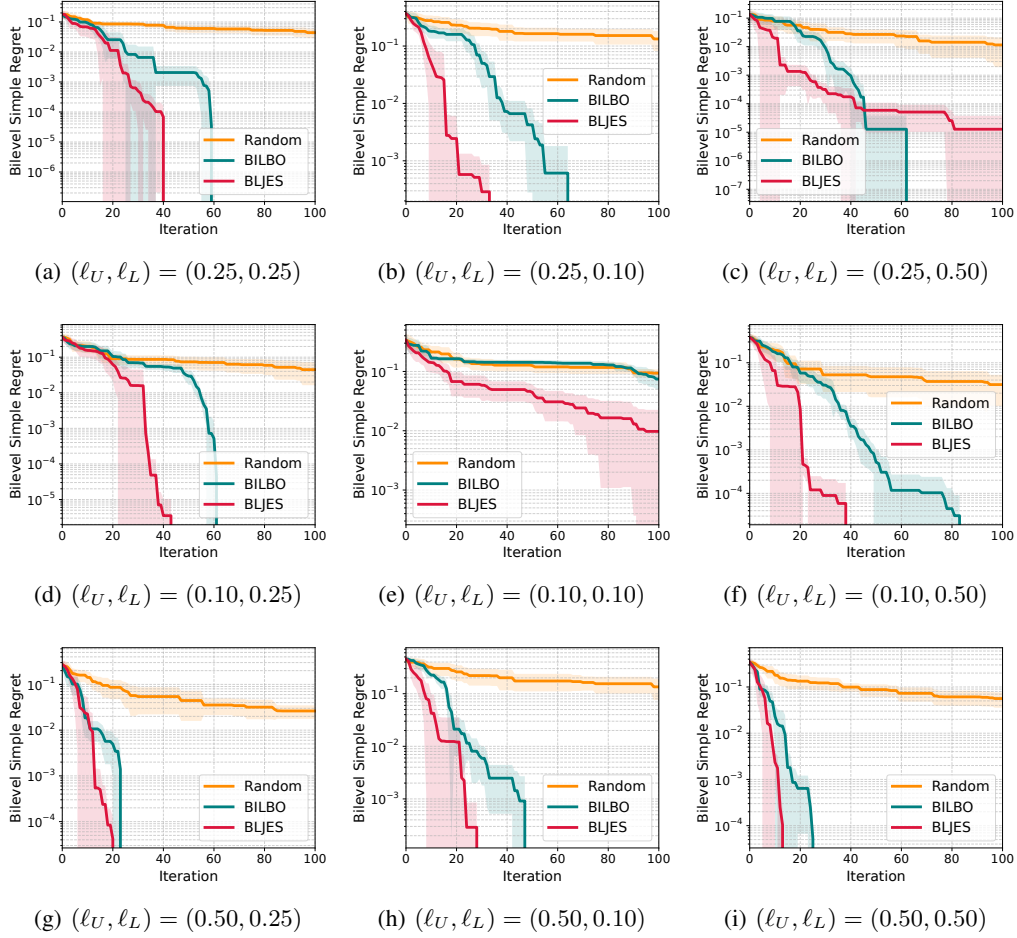


Figure 3: Regret comparison on functions from the GP prior. Since the y-axis is a logarithmic scale, a curve reaching the bottom of the plotting region indicates that the regret has reached zero in all trials.

dimension is 100 for GB and SB (100^2 points), and 10 for Energy and SMD (10^4 points). The last problem is Material data (which is newly created by ourselves), consisting of simulation based evaluations for the crystal structures $\text{Fe}_x\text{Ni}_y\text{Cr}_z$. The problem is to maximize the bulk-modulus $f(\mathbf{x}, \boldsymbol{\theta})$ with respect to $\mathbf{x} = (x, y, z)^\top$ under the minimization constraint of the energy $-g(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ corresponds to the three dimensional position of the atoms ($\boldsymbol{\theta}$ itself is so-called structure descriptor whose dimension is 30). This data have discrete candidates of 54 upper and 500 lower variables ($\mathbf{x}$ and $\boldsymbol{\theta}$), resulting in 27,000 candidates. See Appendix E.1 for further detail.

The results are shown in Fig. 4. BLJES has obviously superior performance except for SMD01 and SMD03. For SMD01 and SMD03, similar performance is shown in BLJES and BILBO, both of which rapidly decrease the regret compared with Random.

5.3 DECOUPLED SETTING

We here evaluate performance on decoupled setting, for which regret comparison is shown in Fig. 5. The objective functions are the same functions used before. We see that MLJES shows smaller regret values for most of problems except only for SMD02 in which BILBO shows better performance. The results indicate that our MI based criterion is effective also for decoupled setting.

5.4 EFFECT OF THE NUMBER OF SAMPLINGS

We evaluate the effect of the number of samplings K on the performance. Figure 6 shows the regret of BLJES with $K = 10, 20, 30$, and 50 on the BG benchmark problem. Note that the result of $K = 30$ is same as Fig. 4 (a). Although $K = 50$ was slightly better than other settings in the end of the optimization, we do not see large differences. Similar tendency has been reported in information-theoretic BO studies [Wang and Jegelka, 2017, Takeno et al., 2022a]. See

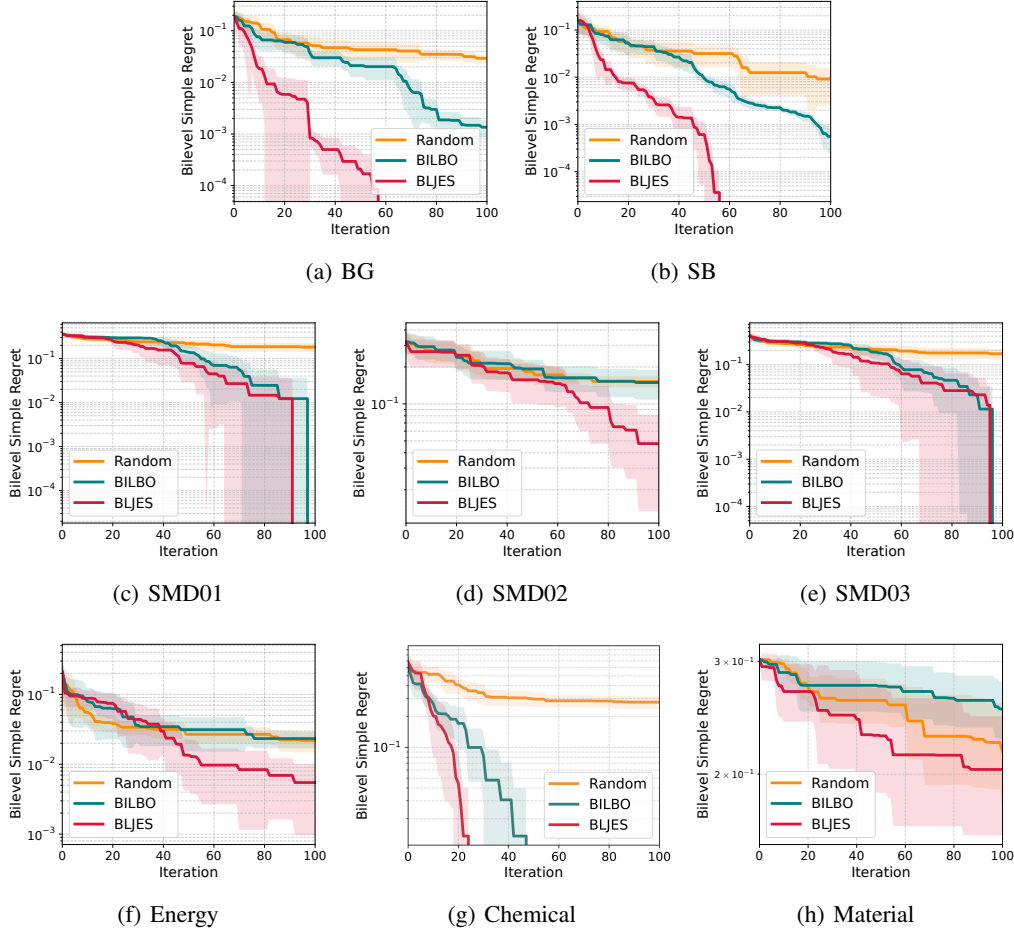


Figure 4: Regret comparison on benchmark and real-world problems.

in Appendix E.5 for the results on other problems.

6 LIMITATIONS

As we described in section 1, the bilevel BO in which both levels are expensive has not been widely studied, and we still have several limitations.

An unresolved theme is the theoretical analysis of the approximation error of MI. The major approximation components of our MI estimator are the lower bound, RFF, and the MC sampling. Providing a theoretical guarantee for the combined approximation error is a challenging issue, though it is actually a common issue for information-theoretic BO methods. For a connection of our lower bound and an independence assumption is discussed in Appendix F, but revealing the approximation error is still an open problem.

Particularly for GP-UCB [Srinivas et al., 2010] based approaches including BILBO, convergence of the regret bound has been widely studied. On the other hand, for information-theoretic BO in general, the regret analysis is also still an

open problem even for the simplest single-level standard problem setting (see Appendix G). Therefore, the regret analysis for information-theoretic BO including BLJES is still needed to be addressed.

7 CONCLUSION

We propose an information-theoretic approach to bilevel Bayesian optimization, called Bilevel optimization via Lower-bound based Joint Entropy Search (BLJES). BLJES considers information gain of optimal points and values of both the upper- and lower-level problems simultaneously, by which we can define a unified criterion that measures the benefit for both the problems. We derive a lower bound based approximation of bilevel information gain, which can be seen as a natural extension of the single level information-theoretic Bayesian optimization. We further propose extensions for decoupled setting and constraint problems. The effectiveness of BLJES is demonstrated through synthetic, benchmark, and real-world functions.

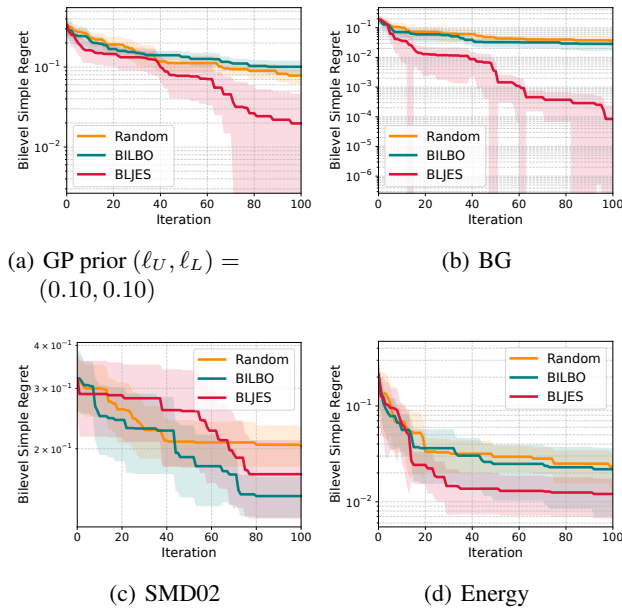


Figure 5: Regret comparison on decoupled setting.

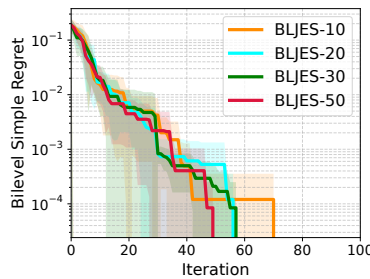


Figure 6: BLJES with different K on the BG benchmark.

ACKNOWLEDGEMENT

This work was partially supported by MEXT KAKENHI (23K21696, 25K03182), MEXT Supporting Pioneering Research through AI for 1,000 Discovery challenges Program (SPReAD) Japan Grant Number JPMXP1726275243, and Data Creation and Utilization Type Material Research and Development Project (Grant No. JP-MXP1122712807) of MEXT.

References

- Malek Abbassi, Abir Chaabani, Lamjed Ben Said, and Nabil Absi. An approximation-based chemical reaction algorithm for combinatorial multi-objective bi-level optimization problems. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 1627–1634, 2021.
- Alireza Aghasi and Saeed Ghadimi. Fully zeroth-order bilevel programming via Gaussian smoothing. *Journal*

of Optimization Theory and Applications, 205(2), Mar 2025.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for Bayesian optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 20577–20612. Curran Associates, Inc., 2023.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. BoTorch: A framework for efficient monte-carlo bayesian optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21524–21538. Curran Associates, Inc., 2020.

Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Max-value entropy search for multi-objective Bayesian optimization. In *Advances in Neural Information Processing Systems 32*, pages 7825–7835. Curran Associates, Inc., 2019.

Ruth Wan Theng Chew, Quoc Phong Nguyen, and Bryan Kian Hsiang Low. BILBO: Bilevel Bayesian optimization. In *Proceedings of the 42th International Conference on Machine Learning*, 2025.

Vedat Dogan and Steven Prestwich. Bilevel optimization by conditional Bayesian optimization. In *Machine Learning, Optimization, and Data Science: 9th International Conference, LOD 2023, Grasmere, UK, September 22–26, 2023, Revised Selected Papers, Part I*, page 243–258, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-53968-8.

Omer Ekmekcioglu, Nursen Aydin, and Juergen Branke. Bayesian optimization of bilevel problems, 2024.

Shi Fu, Fengxiang He, Xinmei Tian, and Dacheng Tao. Convergence of bayesian bilevel optimization. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, pages 16767–16792, 2024.

Miguel González-Duque, Richard Michael, Simon Bartels, Yevgen Zainchkovskyy, Søren Hauberg, and Wouter Boomsma. A survey and benchmark of high-dimensional bayesian optimization of discrete sequences. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc.

Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(57):1809–1837, 2012.

- Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Amar Shah, and Ryan Adams. Predictive entropy search for multi-objective Bayesian optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1492–1501. PMLR, 2016.
- José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems 27*, page 918–926. Curran Associates, Inc., 2014.
- José Miguel Hernández-Lobato, Michael A. Gelbart, Matthew W. Hoffman, Ryan P. Adams, and Zoubin Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *Proceedings of the 32th International Conference on Machine Learning*, volume 37, pages 1699–1707. PMLR, 2015.
- J. Herskovits, A. Leontiev, G. Dias, and G. Santos. Contact shape optimization: a bilevel programming approach. *Structural and Multidisciplinary Optimization*, 20(3):214–221, Nov 2000.
- Matthew W. Hoffman and Zoubin Ghahramani. Output-space predictive entropy search for flexible global optimization. In *NIPS Workshop on Bayesian Optimization*, 2015.
- Carl Hvarfner, Frank Hutter, and Luigi Nardi. Joint entropy search for maximally-informed Bayesian optimization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc.
- Carl Hvarfner, Erik Orm Hellsten, and Luigi Nardi. Vanilla Bayesian optimization performs great in high dimensions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20793–20817. PMLR, 21–27 Jul 2024.
- Masanori Ishikura and Masayuki Karasuyama. Pareto-frontier entropy search with variational lower bound maximization. In *Proceedings of the 42th International Conference on Machine Learning*, 2025.
- Md Monjurul Islam, Hemant Kumar Singh, and Tapabrata Ray. Efficient global optimization for solving computationally expensive bilevel optimization problems. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2018.
- Yang Jiao, Kai Yang, and Chengtao Jian. DTZO: Distributed trilevel zeroth order learning with provable non-asymptotic convergence. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 27873–27916. PMLR, 13–19 Jul 2025.
- Emmanuel Kieffer, Grégoire Danoy, Pascal Bouvry, and Anass Nagih. Bayesian optimization approach of general bi-level problems. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO ’17*, page 1614–1621, New York, NY, USA, 2017. Association for Computing Machinery.
- Johannes Kirschner, Mojmír Mutný, Nicole Hiller, Rasmus Ischebeck, and Andreas Krause. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 3429–3438. PMLR, 2019.
- Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1540–1552. PMLR, 26–28 Aug 2020.
- Mohit Malu, Gautam Dasarathy, and Andreas Spanias. Bayesian optimization in high-dimensional spaces: A brief survey. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–8, 2021.
- Parvin Nazari, Bojian Hou, Davoud Ataee Tarzanagh, Li Shen, and George Michailidis. Stochastic regret guarantees for online zeroth- and first-order bilevel optimization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 147406–147493. Curran Associates, Inc., 2025.
- Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and multidisciplinary optimization*, 48(3):607–626, 2013.
- Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5171–5180. PMLR, 2019.
- Yuyang Qiu, Uday Shanbhag, and Farzad Yousefian. Zeroth-order methods for nondifferentiable, nonconvex, and hierarchical federated optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*,

- volume 36, pages 3425–3438. Curran Associates, Inc., 2023.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.
- Maria Laura Santoni, Elena Raponi, Renato De Leone, and Carola Doerr. Comparison of high-dimensional Bayesian optimization algorithms on bbo. *ACM Transactions on Evolutionary Learning*, 4(3), July 2024.
- K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B*, 89:205118, May 2014.
- Amar Shah and Zoubin Ghahramani. Parallel predictive entropy search for batch global optimization of expensive objective functions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, page 3330–3338. Curran Associates, Inc., 2015.
- Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. Test problem construction for single-objective bilevel optimization. *Evolutionary Computation*, 22(3):439–477, 2014.
- Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022. Omnipress, 2010.
- Varun Suryan, Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. Handling inverse optimal control problems using evolutionary bilevel optimization. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1893–1900, 2016.
- Shinya Suzuki, Shion Takeno, Tomoyuki Tamura, Kazuki Shitara, and Masayuki Karasuyama. Multi-objective Bayesian optimization using Pareto-frontier entropy. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 9279–9288. PMLR, 2020.
- Shion Takeno, Hitoshi Fukuoka, Yuhki Tsukada, Toshiyuki Koyama, Motoki Shiga, Ichiro Takeuchi, and Masayuki Karasuyama. A generalized framework of multi-fidelity max-value entropy search through joint entropy. *Neural Computation*, 34(10):2145–2203, 2022a.
- Shion Takeno, Tomoyuki Tamura, Kazuki Shitara, and Masayuki Karasuyama. Sequential and parallel constrained max-value entropy search via information lower bound. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20960–20986. PMLR, 2022b.
- Shion Takeno, Yu Inatsu, Masayuki Karasuyama, and Ichiro Takeuchi. Posterior sampling-based bayesian optimization with tighter Bayesian regret bounds. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org, 2024.
- Ben Tu, Axel Gandy, Nikolas Kantas, and Behrang Shafei. Joint entropy search for multi-objective Bayesian optimization. *Advances in Neural Information Processing Systems*, 35:9922–9938, 2022.
- Bing Wang, Hemant Kumar Singh, and Tapabrata Ray. Comparing expected improvement and kriging believer for expensive bilevel optimization. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 1635–1642, 2021.
- Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3627–3635. PMLR, 2017.
- Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Freitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55(1):361–387, January 2016.
- Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024.

Information-Theoretic Bayesian Optimization for Bilevel Optimization Problems (Supplementary Material)

Takuya Kanayama¹

Yuki Ito¹

Tomoyuki Tamura²

Masayuki Karasuyama¹

¹Dept. of Computer Science, Nagoya Institute of Technology, Nagoya, Aichi, Japan

²Dept. of Physical Science and Engineering, Nagoya Institute of Technology, Nagoya, Aichi, Japan

A DERIVATION OF LOWER BOUND

A.1 PROOF OF THEOREM 3.1

From Bayes theorem,

$$p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))} \leq f^*, \mathcal{D}_t^+) = \frac{p(f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^* \mid y_{(\mathbf{x}, \boldsymbol{\theta})}^f, \mathcal{D}_t^+) p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t^+)}{p(f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^* \mid \mathcal{D}_t^+)}. \quad (10)$$

All the three densities in the right hand side, the analytical representations can be derived as follows.

- The probability $p(f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^* \mid y_{(\mathbf{x}, \boldsymbol{\theta})}^f, \mathcal{D}_t^+)$ is calculated by the density

$$f_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))} \mid y_{(\mathbf{x}, \boldsymbol{\theta})}^f, \mathcal{D}_t^+ \sim \mathcal{N}(m_1^f, \{s_1^f\}^2),$$

for which the mean m_1^f and variance $\{s_1^f\}^2$ can be derived by considering the conditional density of the joint posterior of $f_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}$, $y_{(\mathbf{x}, \boldsymbol{\theta})}^f$, and $f_{(\mathbf{x}^*, \boldsymbol{\theta}^*)}$ as

$$\begin{aligned} m_1^f &= \mu_t^f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \\ &+ \begin{bmatrix} \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}, \boldsymbol{\theta})) \\ \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \end{bmatrix}^\top \begin{bmatrix} \{\sigma_t^f(\mathbf{x}, \boldsymbol{\theta})\}^2 + \{\sigma_{\text{noise}}^f\}^2 & \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \\ \text{Cov}_t^f((\mathbf{x}^*, \boldsymbol{\theta}^*), (\mathbf{x}, \boldsymbol{\theta})) & \{\sigma_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2 \end{bmatrix}^{-1} \begin{bmatrix} y_{(\mathbf{x}, \boldsymbol{\theta})}^f - \mu_t^f(\mathbf{x}, \boldsymbol{\theta}) \\ f_{(\mathbf{x}^*, \boldsymbol{\theta}^*)} - \mu_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*) \end{bmatrix} \\ \{s_1^f\}^2 &= \{\sigma_t^f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))\}^2 \\ &- \begin{bmatrix} \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}, \boldsymbol{\theta})) \\ \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \end{bmatrix}^\top \begin{bmatrix} \{\sigma_t^f(\mathbf{x}, \boldsymbol{\theta})\}^2 + \{\sigma_{\text{noise}}^f\}^2 & \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \\ \text{Cov}_t^f((\mathbf{x}^*, \boldsymbol{\theta}^*), (\mathbf{x}, \boldsymbol{\theta})) & \{\sigma_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2 \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}, \boldsymbol{\theta})) \\ \text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \end{bmatrix}. \end{aligned}$$

Note that $\text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'))$ is the posterior covariance between $f(\mathbf{x}, \boldsymbol{\theta})$ and $f(\mathbf{x}', \boldsymbol{\theta}')$, given $\mathcal{D}_t$. By using m_1^f and s_1^f , we have

$$p(f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^* \mid y_{(\mathbf{x}, \boldsymbol{\theta})}^f, \mathcal{D}_t^+) = \begin{cases} \Phi\left(\frac{f^* - m_1^f}{s_1^f}\right) & \text{if } \mathbf{x} \neq \mathbf{x}^*, \\ 1 & \text{otherwise,} \end{cases} \quad (11)$$

where Φ is the cumulative density function (CDF) of the standard normal distribution.

- Next, to calculate the denominator $p(f_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))} \leq f^* \mid \mathcal{D}_t^+)$, we consider the density

$$f_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))} \mid \mathcal{D}_t^+ \sim \mathcal{N}(m_2^f, \{s_2^f\}^2),$$

for which the mean m_2^f and variance $\{s_2^f\}^2$ can be derived by considering the conditional density of the joint posterior of $f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))$ and $f(\mathbf{x}^*, \boldsymbol{\theta}^*)$ as

$$m_2^f = \mu_t^f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) + \frac{\text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}^*, \boldsymbol{\theta}^*))}{\{\sigma_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2} (f(\mathbf{x}^*, \boldsymbol{\theta}^*) - \mu_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*))$$

$$\{s_2^f\}^2 = \sigma_t^f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) - \frac{\{\text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}^*, \boldsymbol{\theta}^*))\}^2}{\{\sigma_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2}$$

Then, we obtain

$$p(f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^* \mid \mathcal{D}_t^+) = \begin{cases} \Phi\left(\frac{f^* - m_2^f}{s_2^f}\right) & \text{if } \mathbf{x} \neq \mathbf{x}^* \\ 1 & \text{otherwise} \end{cases} \quad (12)$$

- The density $p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t^+)$ can also be derived by a similar approach as

$$y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t^+ \sim \mathcal{N}(m_3^f, s_3^f),$$

where

$$m_3^f = \mu_t^f(\mathbf{x}, \boldsymbol{\theta}) + \frac{\text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*))}{\{\sigma_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2} (f(\mathbf{x}^*, \boldsymbol{\theta}^*) - \mu_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*))$$

$$\{s_3^f\}^2 = \{\sigma_t^f(\mathbf{x}, \boldsymbol{\theta})\}^2 + \{\sigma_{\text{noise}}^f\}^2 - \frac{\{\text{Cov}_t^f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*))\}^2}{\{\sigma_t^f(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2}.$$

Therefore,

$$p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t^+) = \phi\left(\frac{y_{(\mathbf{x}, \boldsymbol{\theta})}^f - m_3^f}{s_3^f}\right) / s_3^f. \quad (13)$$

where ϕ is the density function of the standard normal distribution.

By substituting (11), (12), and (13) into (10), we obtain (6).

A.2 ANALYTICAL REPRESENTATION OF $p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+)$

In the case of $p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+)$, almost the same derivation can be applied as (6). Therefore, we here only show the final result

$$p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+) = \begin{cases} \Phi\left(\frac{g^* - m_1^g}{s_1^g}\right) \phi\left(\frac{y_{(\mathbf{x}, \boldsymbol{\theta})}^g - m_3^g}{s_3^g}\right) / \Phi\left(\frac{g^* - m_2^g}{s_2^g}\right) s_3^g & \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \\ \phi\left(\frac{y_{(\mathbf{x}, \boldsymbol{\theta})}^g - m_3^g}{s_3^g}\right) / s_3^g & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned}
m_1^g &= \mu_t^g(\mathbf{x}^*, \boldsymbol{\theta}) + \begin{bmatrix} \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}, \boldsymbol{\theta})) \\ \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \end{bmatrix}^\top \begin{bmatrix} \{\sigma_t^g(\mathbf{x}, \boldsymbol{\theta})\}^2 + \{\sigma_{\text{noise}}^g\}^2 & \text{Cov}_t^g((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \\ \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}^*), (\mathbf{x}, \boldsymbol{\theta})) & \{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2 \end{bmatrix}^{-1} \begin{bmatrix} y_{(\mathbf{x}, \boldsymbol{\theta})}^g - \mu_t^g(\mathbf{x}, \boldsymbol{\theta}) \\ g(\mathbf{x}^*, \boldsymbol{\theta}^*) - \mu_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*) \end{bmatrix}, \\
\{s_1^g\}^2 &= \{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta})\}^2 \\
&\quad - \begin{bmatrix} \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}, \boldsymbol{\theta})) \\ \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \end{bmatrix}^\top \begin{bmatrix} \{\sigma_t^g(\mathbf{x}, \boldsymbol{\theta})\}^2 + \{\sigma_{\text{noise}}^g\}^2 & \text{Cov}_t^g((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \\ \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}^*), (\mathbf{x}, \boldsymbol{\theta})) & \{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2 \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}, \boldsymbol{\theta})) \\ \text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*)) \end{bmatrix}, \\
m_2^g &= \mu_t^g(\mathbf{x}^*, \boldsymbol{\theta}) + \frac{\text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*))}{\{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2} (g(\mathbf{x}^*, \boldsymbol{\theta}^*) - \mu_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)) \\
\{s_2^g\}^2 &= \{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta})\}^2 - \frac{\{\text{Cov}_t^g((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*))\}^2}{\{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2}, \\
m_3^g &= \mu_t^g(\mathbf{x}, \boldsymbol{\theta}) + \frac{\text{Cov}_t^g((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*))}{\{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2} (g(\mathbf{x}^*, \boldsymbol{\theta}^*) - \mu_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)), \\
\{s_3^g\}^2 &= \{\sigma_t^g(\mathbf{x}, \boldsymbol{\theta})\}^2 + \{\sigma_{\text{noise}}^g\}^2 - \frac{\{\text{Cov}_t^g((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}^*, \boldsymbol{\theta}^*))\}^2}{\{\sigma_t^g(\mathbf{x}^*, \boldsymbol{\theta}^*)\}^2},
\end{aligned}$$

and $\text{Cov}_t^g((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'))$ is the posterior covariance between $g(\mathbf{x}, \boldsymbol{\theta})$ and $g(\mathbf{x}', \boldsymbol{\theta}')$ given $\mathcal{D}_t$.

B DETAIL OF GRADIENT COMPUTATIONS

First, we consider the gradient for $\partial \tilde{f}(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))/\partial \mathbf{x}$, which is required to obtain the sample of $\mathbf{x}^*, \boldsymbol{\theta}^*, f^*$, and g^* . For $\tilde{\boldsymbol{\theta}}^*(\mathbf{x}) = \arg \max_{\boldsymbol{\theta}} \tilde{g}(\mathbf{x}, \boldsymbol{\theta})$, the implicit function theorem derives

$$\frac{\partial \tilde{\boldsymbol{\theta}}^*(\mathbf{x})}{\partial \mathbf{x}^\top} = - \left\{ \frac{\partial^2 \tilde{g}(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\mathbf{x})} \right\}^{-1} \frac{\partial^2 \tilde{g}(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \mathbf{x}^\top} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\mathbf{x})},$$

from which we can calculate

$$\frac{\partial \tilde{f}(\mathbf{x}, \tilde{\boldsymbol{\theta}}^*(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial \tilde{f}(\mathbf{x}, \boldsymbol{\theta})}{\partial \mathbf{x}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^*(\mathbf{x})} + \left\{ \frac{\partial \tilde{\boldsymbol{\theta}}^*(\mathbf{x})}{\partial \mathbf{x}^\top} \right\}^\top \frac{\partial \tilde{f}(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^*(\mathbf{x})}.$$

Next, we consider the acquisition function maximization. Let

$$\tilde{a}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\theta}') := \log \frac{p(\tilde{y}_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \tilde{f}(\mathbf{x}, \boldsymbol{\theta}') \leq \tilde{f}^*, \tilde{\mathcal{D}}_t^+)}{p(\tilde{y}_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t)} + \log \frac{p(\tilde{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \tilde{g}(\tilde{\mathbf{x}}^*, \boldsymbol{\theta}) \leq \tilde{g}^*, \tilde{\mathcal{D}}_t^+)}{p(\tilde{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t)}$$

be the inside of the expectation of (5) in which $\boldsymbol{\theta}^*(\mathbf{x})$ is replaced with $\boldsymbol{\theta}'$, and variables in Ω is replaced by a sample, denoted with $\tilde{\cdot}$. Note that $\tilde{\mathcal{D}}_t^+ = \mathcal{D}_t \cup (\tilde{\mathbf{x}}^*, \tilde{\boldsymbol{\theta}}^*, \tilde{f}^*, \tilde{g}^*)$. Then, the gradient with respect to $\mathbf{x}$ can be written as

$$\frac{\partial \tilde{a}(\mathbf{x}, \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^*(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial \tilde{a}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\theta}')}{\partial \mathbf{x}} \bigg|_{\boldsymbol{\theta}'=\tilde{\boldsymbol{\theta}}^*(\mathbf{x})} + \left\{ \frac{\partial \tilde{\boldsymbol{\theta}}^*(\mathbf{x})}{\partial \mathbf{x}^\top} \right\}^\top \frac{\partial \tilde{a}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\theta}')}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}'=\tilde{\boldsymbol{\theta}}^*(\mathbf{x})}$$

The gradient with respect $\boldsymbol{\theta}$ can be obtained through the usual derivative.

C LOWER BOUND OF DECOUPLED SETTING

The lower bound of the information gain for the upper-level observation is

$$\begin{aligned}
\text{MI}(y_{(\mathbf{x}, \boldsymbol{\theta})}^f ; f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^* \mid \mathcal{D}_t) &= \mathbb{E}_{\Omega} \left[\log \frac{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t)} \right] \\
&= \mathbb{E}_{f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*} \left[\mathbb{E}_{y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t} \left[\log \frac{q(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t)} \right] \right. \\
&\quad \left. + \text{KL} \left(p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \parallel q(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \right) \right] \\
&\geq \mathbb{E}_{\Omega} \left[\log \frac{q(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t)} \right] =: \text{LB}^f(\mathbf{x}, \boldsymbol{\theta}).
\end{aligned}$$

By setting the variational distribution as

$$\begin{aligned}
q(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) &:= p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*, g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+) \\
&= p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*, \mathcal{D}_t^+),
\end{aligned}$$

we obtain the lower bound (7).

D EXTENSION FOR CONSTRAINT PROBLEMS

Let

$$\begin{aligned}
\mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta})}^f &:= (f(\mathbf{x}, \boldsymbol{\theta}), c_1^U(\mathbf{x}, \boldsymbol{\theta}), \dots, c_N^U(\mathbf{x}, \boldsymbol{\theta}))^\top, \\
\mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta})}^g &:= (g(\mathbf{x}, \boldsymbol{\theta}), c_1^L(\mathbf{x}, \boldsymbol{\theta}), \dots, c_M^L(\mathbf{x}, \boldsymbol{\theta}))^\top,
\end{aligned}$$

be the vectors in which the objective function and the constraint functions are concatenated for the upper- and the lower-level problems, respectively, and

$$\begin{aligned}
\mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^f &:= (y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_1^U}, \dots, y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_N^U})^\top, \\
\mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g &:= (y_{(\mathbf{x}, \boldsymbol{\theta})}^g, y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_1^L}, \dots, y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_M^L})^\top,
\end{aligned}$$

are the counterparts of noisy observations, where $y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_n^L} := c_n^L(\mathbf{x}, \boldsymbol{\theta}) + \epsilon_n^L$, $\epsilon_n^L \sim \mathcal{N}(0, \{\sigma_{\text{noise}}^{c_n^L}\}^2)$ and $y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_m^U} := c_m^U(\mathbf{x}, \boldsymbol{\theta}) + \epsilon_m^U$, $\epsilon_m^U \sim \mathcal{N}(0, \{\sigma_{\text{noise}}^{c_m^U}\}^2)$. We observe $(\mathbf{y}_{\mathbf{x}, \boldsymbol{\theta}}^f, \mathbf{y}_{\mathbf{x}, \boldsymbol{\theta}}^g)$ at every BO iteration for selected $(\mathbf{x}, \boldsymbol{\theta})$, i.e., $\mathcal{D}_t = \{(\mathbf{x}_i, \boldsymbol{\theta}_i, \mathbf{y}_{\mathbf{x}_i, \boldsymbol{\theta}_i}^f, \mathbf{y}_{\mathbf{x}_i, \boldsymbol{\theta}_i}^g)\}_{i=1}^n$ in the constraint setting. In addition to f and g , the independent GPs are also fitted to c_n^U and c_m^L , for which the posteriors given $\mathcal{D}_t$ are written as $\mathcal{N}(\mu_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta}), \{\sigma_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta})\}^2)$ and $\mathcal{N}(\mu_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta}), \{\sigma_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta})\}^2)$, respectively.

D.1 LOWER BOUND

The MI and its lower bound can be derived by the same approach as (3):

$$\begin{aligned}
\text{MI}(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g; f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^* | \mathcal{D}_t) &= \mathbb{E}_\Omega \left[\log \frac{p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)}{p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | \mathcal{D}_t)} \right] \\
&= \mathbb{E}_{f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*} \left[\mathbb{E}_{\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t} \left[\log \frac{q(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)}{p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | \mathcal{D}_t)} \right] \right. \\
&\quad \left. + \text{KL} \left(p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \parallel q(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \right) \right] \\
&\geq \mathbb{E}_\Omega \left[\log \frac{q(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)}{p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | \mathcal{D}_t)} \right] =: \text{LB}_c(\mathbf{x}, \boldsymbol{\theta}),
\end{aligned}$$

where here $\mathcal{D}_t^+ := \mathcal{D}_t \cup \{(\mathbf{x}^*, \boldsymbol{\theta}^*, f^*, g^*)\}$.

To define the variational distribution q , we follow the same approach as information-theoretic constraint BO proposed by [Takeno et al., 2022b]. Let $\mathcal{A}^f = \{(c_0, \mathbf{c}) \mid c_0 \geq f^*, \mathbf{c} \geq 0, c_0 \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^N\}$ and $\mathcal{A}^g = \{(c_0, \mathbf{c}) \mid c_0 \geq g^*, \mathbf{c} \geq 0, c_0 \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^M\}$. When f^* is given, $\mathcal{A}^f$ is the region that $\mathbf{h}_{\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})}^f$ cannot exist for $\forall \mathbf{x}$. When g^* is given, $\mathcal{A}^g$ is the region that $\mathbf{h}_{\mathbf{x}^*, \boldsymbol{\theta}}^g$ cannot exist for $\forall \boldsymbol{\theta}$. Based on the same simplification of the conditioning discussed in section 3.1, we define the variational distribution as

$$q(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) := p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | \mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f, \mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g, \mathcal{D}_t^+),$$

where $\bar{\mathcal{A}}^f$ and $\bar{\mathcal{A}}^g$ are the complement sets of $\mathcal{A}^f$ and $\mathcal{A}^g$, respectively. As a result, we see

$$\begin{aligned}
\text{LB}_c(\mathbf{x}, \boldsymbol{\theta}) &= \mathbb{E}_\Omega \left[\log \frac{p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | \mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f, \mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g, \mathcal{D}_t^+)}{p(\mathbf{y}_{(x,\theta)}^f, \mathbf{y}_{(x,\theta)}^g | \mathcal{D}_t)} \right] \\
&= \mathbb{E}_\Omega \left[\log \frac{p(\mathbf{y}_{(x,\theta)}^f | \mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f, \mathcal{D}_t^+)}{p(\mathbf{y}_{(x,\theta)}^f | \mathcal{D}_t)} + \log \frac{p(\mathbf{y}_{(x,\theta)}^g | \mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g, \mathcal{D}_t^+)}{p(\mathbf{y}_{(x,\theta)}^g | \mathcal{D}_t)} \right].
\end{aligned}$$

D.2 ANALYTICAL REPRESENTATION OF VARIATIONAL DISTRIBUTION

From Bayes theorem,

$$p(\mathbf{y}_{(x,\theta)}^f | \mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f, \mathcal{D}_t^+) = \frac{p(\mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f | \mathbf{y}_{(x,\theta)}^f, \mathcal{D}_t^+) p(\mathbf{y}_{(x,\theta)}^f | \mathcal{D}_t^+)}{p(\mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f | \mathcal{D}_t^+)}. \quad (14)$$

The density $p(\mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f | \mathbf{y}_{(x,\theta)}^f, \mathcal{D}_t^+)$ is an $(N+1)$ -dimensional independent Gaussian distribution, for which the first dimension is $\mathcal{N}(m_1^f, \{s_1^f\}^2)$ shown in Appendix A.1 and from the second to the $(N+1)$ -th dimension is $\mathcal{N}(m^{c_n^U}, \{s^{c_n^U}\}^2)$ where

$$\begin{aligned}
m^{c_n^U} &= \mu_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) + \frac{\text{Cov}_t^{c_n^U}((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}, \boldsymbol{\theta}))}{\{\sigma_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta})\}^2} (y_{(x,\theta)}^{c_n^U} - \mu_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta})), \\
\{s^{c_n^U}\}^2 &= \sigma_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) - \frac{\{\text{Cov}_t^{c_n^U}((\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})), (\mathbf{x}, \boldsymbol{\theta}))\}^2}{\{\sigma_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta})\}^2}.
\end{aligned}$$

As a result, we can derive

$$\begin{aligned}
p(\mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f \mid \mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^f, \mathcal{D}_t^+) &= 1 - (1 - \Phi(\frac{f^* - m_1^f}{s_1^f})) \prod_{n=1}^N (1 - \Phi(\frac{0 - m_n^{c_n^U}}{s_n^{c_n^U}})), \\
p(\mathbf{h}_{(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}^f \in \bar{\mathcal{A}}^f \mid \mathcal{D}_t^+) &= 1 - (1 - \Phi(\frac{f^* - m_2^f}{s_2^f})) \prod_{n=1}^N (1 - \Phi(\frac{0 - \mu_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))}{\sigma_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))})), \\
p(\mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t^+) &= \phi(\frac{y_{(\mathbf{x}, \boldsymbol{\theta})}^f - m_3^f}{s_3^f}) \prod_{n=1}^N \phi(\frac{y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_n^U} - \mu_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta})}{\sigma_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta})}) / (s_3^f \sigma_t^{c_n^U}(\mathbf{x}, \boldsymbol{\theta})).
\end{aligned}$$

Similarly, for the lower-level density, Bayes theorem transforms

$$p(\mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g, \mathcal{D}_t^+) = \frac{p(\mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g \mid \mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g, \mathcal{D}_t^+) p(\mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t^+)}{p(\mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g \mid \mathcal{D}_t^+)}$$

Here again, the density of the first dimension of $p(\mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \mid \mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g, \mathcal{D}_t^+)$ is $\mathcal{N}(m_1^g, \{s_1^g\}^2)$ shown in Appendix A.1 and from the second to the $(M + 1)$ -th dimension is $\mathcal{N}(m_m^{c_m^L}, \{s_m^{c_m^L}\}^2)$ where

$$\begin{aligned}
m_m^{c_m^L} &= \mu_t^{c_m^L}(\mathbf{x}^*, \boldsymbol{\theta}) + \frac{\text{Cov}_t^{c_m^L}((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}, \boldsymbol{\theta}))}{\{\sigma_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta})\}^2} (y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_m^L} - \mu_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta})), \\
\{s_m^{c_m^L}\}^2 &= \sigma_t^{c_m^L}(\mathbf{x}^*, \boldsymbol{\theta}) - \frac{\{\text{Cov}_t^{c_m^L}((\mathbf{x}^*, \boldsymbol{\theta}), (\mathbf{x}, \boldsymbol{\theta}))\}^2}{\{\sigma_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta})\}^2}.
\end{aligned}$$

As a result

$$\begin{aligned}
p(\mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g \mid \mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g, \mathcal{D}_t^+) &= 1 - \left(1 - \Phi\left(\frac{g^* - m_1^g}{s_1^g}\right)\right) \prod_{m=1}^M \left(1 - \Phi\left(\frac{0 - m_m^{c_m^L}}{s_m^{c_m^L}}\right)\right) \\
p(\mathbf{h}_{(\mathbf{x}^*, \boldsymbol{\theta})}^g \in \bar{\mathcal{A}}^g \mid \mathcal{D}_t^+) &= 1 - \left(1 - \Phi\left(\frac{g^* - m_2^g}{s_2^g}\right)\right) \prod_{m=1}^M \left(1 - \Phi\left(\frac{0 - \mu_t^{c_m^L}(\mathbf{x}^*, \boldsymbol{\theta})}{\sigma_t^{c_m^L}(\mathbf{x}^*, \boldsymbol{\theta})}\right)\right) \\
p(\mathbf{y}_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t^+) &= \phi\left(\frac{y_{(\mathbf{x}, \boldsymbol{\theta})}^g - m_3^g}{s_3^g}\right) \prod_{m=1}^M \phi\left(\frac{y_{(\mathbf{x}, \boldsymbol{\theta})}^{c_m^L} - \mu_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta})}{\sigma_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta})}\right) / (s_3^g \sigma_t^{c_m^L}(\mathbf{x}, \boldsymbol{\theta}))
\end{aligned}$$

E SUPPLEMENTARY FOR EXPERIMENTS

E.1 OTHER DETAILS OF EXPERIMENTAL SETTINGS

We used the `SingleTaskGP` model of BoTorch [Balandat et al., 2020] to define the GPs. The output of each benchmark function are transformed by signed log1p function $\text{sign}(y) \log(1 + |y|)$ except for BraninHoo and Goldstein-price for which the transformation shown by [Picheny et al., 2013] was used. For benchmark functions, the input space is scaled to $[0, 1]^{d_{\mathcal{X}} + d_{\Theta}}$ from the original input domain. For the chemical dataset, we re-defined the true objective function values as $f(\mathbf{x}, \boldsymbol{\theta}) = -\log((\max_{\mathbf{x}', \boldsymbol{\theta}'} f_{\text{ori}}(\mathbf{x}', \boldsymbol{\theta}') + 10^{-4}) - f_{\text{ori}}(\mathbf{x}, \boldsymbol{\theta}))$ and $g(\mathbf{x}, \boldsymbol{\theta}) = -\log((\max_{\mathbf{x}', \boldsymbol{\theta}'} g_{\text{ori}}(\mathbf{x}', \boldsymbol{\theta}') + 10^{-4}) - g_{\text{ori}}(\mathbf{x}, \boldsymbol{\theta}))$, where f_{ori} and g_{ori} are original functions in the dataset. This is a log transformation applied to the difference from the maximum value (instead of 0). We employed this transformation because, in this dataset, many values are concentrated on the small scale differences around $\max f_{\text{ori}}(\mathbf{x}, \boldsymbol{\theta})$ and $\max g_{\text{ori}}(\mathbf{x}, \boldsymbol{\theta})$, respectively.

In the Material data, the bcc (body-centered cubic) structure $\text{Fe}_x\text{Ni}_y\text{Cr}_z$ with the constraints $x + y + z = 54$ and $x, y, z \geq 15$ is considered. For each candidate $(\mathbf{x}, \boldsymbol{\theta})$, a classical interatomic potential based structure relaxation is performed to obtain the energy and the bulk-modulus. For a given $\mathbf{x}$, although the bcc structure specifies the coordinates at which atoms are

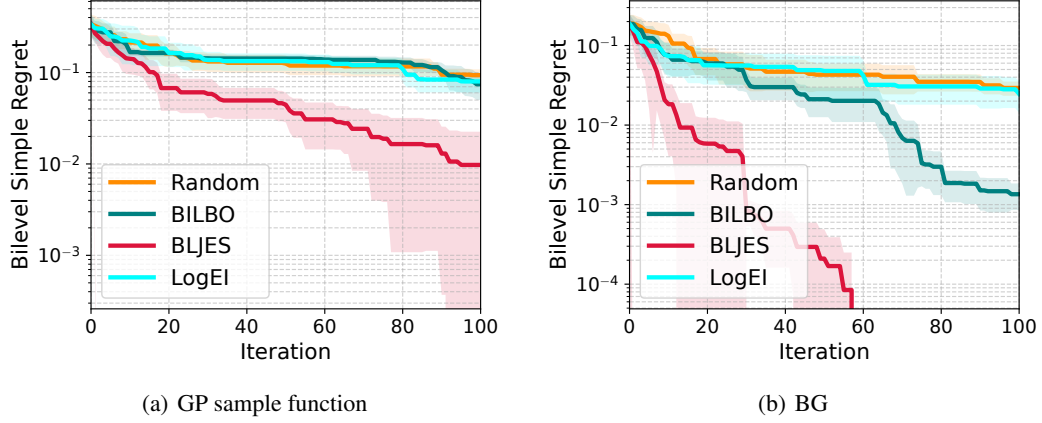


Figure 7: Additional comparison with logEI (from BOTorch) applied to the upper objective function $f(\mathbf{x}, \boldsymbol{\theta})$. Matern kernel ($\nu = 2.5$) is used. The length-scale, the output scale, and the noise variance are optimized by the marginal likelihood at every iteration. Other settings are the same as the original experiments.

located, there remains freedom regarding which atom (Fe or Ni or Cr) occupies each site; we randomly generated 500 patterns of atom assignments for each of $\mathbf{x}$. For $\boldsymbol{\theta}$, we employed a histogram of atomic distances (we created 20 bins in $[1, 6]$ Å), also called the radial distribution function (RDF), which is a well-known structural descriptor [e.g., Schütt et al., 2014]. RDF can be defined for each pair of atoms, and after removing redundant dimensions (dimensions with all zero values), we obtain 30 dimensional $\boldsymbol{\theta}$. Since this data has a high dimensional input, we employ the LogNormal ($\mathcal{LN}$) hyper-prior on the Gaussian kernel length scale ℓ , which follows [Hvarfner et al., 2024]:

$$\ell \sim \mathcal{LN}(\mu_0 + \log(d_{\mathcal{X}} + d_{\Theta})/2, \sigma_0),$$

where $\mu_0 = \sqrt{2}$ and $\sigma_0 = \sqrt{3}$ that are suggestions by [Hvarfner et al., 2024]. This is a common setting for BILBO and BLJES.

E.2 COMPARISON WITH STANDARD SINGLE-LEVEL EI

As an additional naïve baseline, we here apply the standard single-level expected improvement (EI) acquisition function to the GP sample paths ($\ell_U = 0.01$ and $\ell_L = 0.10$) and BG functions. Since EI cannot deal with the constraint of lower-level optimality, we simply ignore the lower-level problem (i.e., calculating EI of $f(\mathbf{x}, \boldsymbol{\theta})$). We used the BOTorch implementation of the logEI [Ament et al., 2023] acquisition function. The results shown in Fig. 7 indicate the performance of the naïve application of EI is similar to the random selection. Therefore, this confirms that the naïve application of the single level acquisition function can cause a critical failure of the optimization.

E.3 LARGER NOISE SETTING

Figure 8 shows results with a stronger noise setting (the noise standard deviation is set as 10^{-1}). We do not see large difference for the relative performance among compared methods compared with the small noise setting shown in Fig. 4.

E.4 ADDITIONAL RESULTS ON DECOUPLED SETTING

Figure 9 shows regret in decoupled setting for all different length scale settings of the GP prior, and Fig. 10 shows the results on other benchmark functions. We obviously see that BLJES was superior or comparable to BILBO.

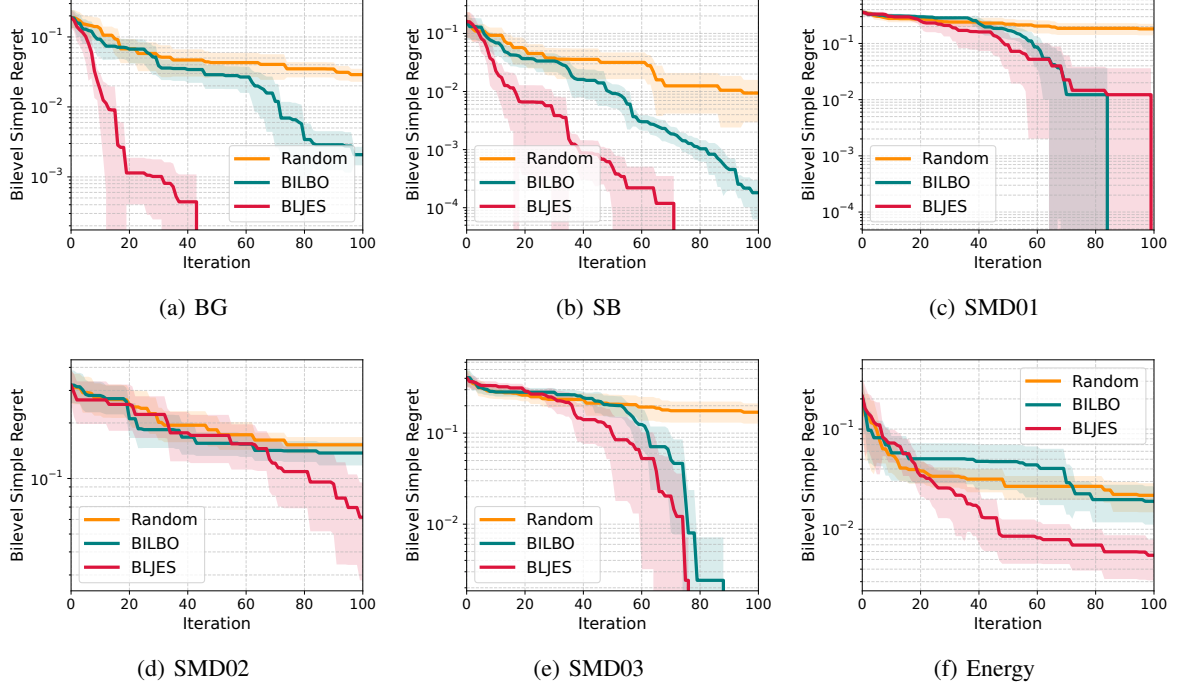


Figure 8: Regret comparison with 10^{-1} noise standard deviation.

E.5 ADDITIONAL RESULTS ON EFFECT OF THE NUMBER OF SAMPLINGS

Figure 11 shows results of BLJES for different K . We do not see particularly large differences among different K settings in these benchmarks.

E.6 CONTINUOUS DOMAIN

Figure 12 shows the regret in the case of $\mathcal{X}$ and Θ are the continuous space. We employed gradient based optimizers for both of the bilevel problem defined by sample paths and the acquisition function maximization (gradient of a bilevel problem is discussed in Appendix B). Here, BILBO is not performed because Chew et al. [2025] only discuss the finite domain. We see that BLJES efficiently decreases the regret even in the continuous space. Only in SMD02, BLJES was not efficient compared with the random selection.

E.7 CONSTRAINT PROBLEMS

For empirical evaluation, we employed problems from the bilevel optimization benchmark [Sinha et al., 2014], denoted as SDM09 ($d_{\mathcal{X}} = 2, d_{\Theta} = 2, N = 1, M = 1$), 10 ($d_{\mathcal{X}} = 2, d_{\Theta} = 2, N = 2, M = 1$), 11 ($d_{\mathcal{X}} = 2, d_{\Theta} = 2, N = 1, M = 1$), and 12 ($d_{\mathcal{X}} = 2, d_{\Theta} = 2, N = 3, M = 2$). The number of grid points in each dimension is 10 (10^4 points). The evaluation metric is

$$\min_{i \in [n_0+t]} \max_{h \in \{f, g, c_1^U, \dots, c_N^U, c_1^L, \dots, c_M^L\}} r_h(\mathbf{x}_i, \boldsymbol{\theta}_i)$$

where r_h become r_f and r_g shown in section 5 if $h = f$ or g , and

$$r_c(\mathbf{x}_i, \boldsymbol{\theta}_i) = \max(0, -c(\mathbf{x}_i, \boldsymbol{\theta}_i)) / \max_{\mathbf{x}, \boldsymbol{\theta}}(\max(0, -c(\mathbf{x}, \boldsymbol{\theta}))), c \in \{c_1^U, \dots, c_N^U, c_1^L, \dots, c_M^L\}$$

if $h \in \{c_1^U, \dots, c_N^U, c_1^L, \dots, c_M^L\}$. The other settings are same as described in the beginning of section 5.

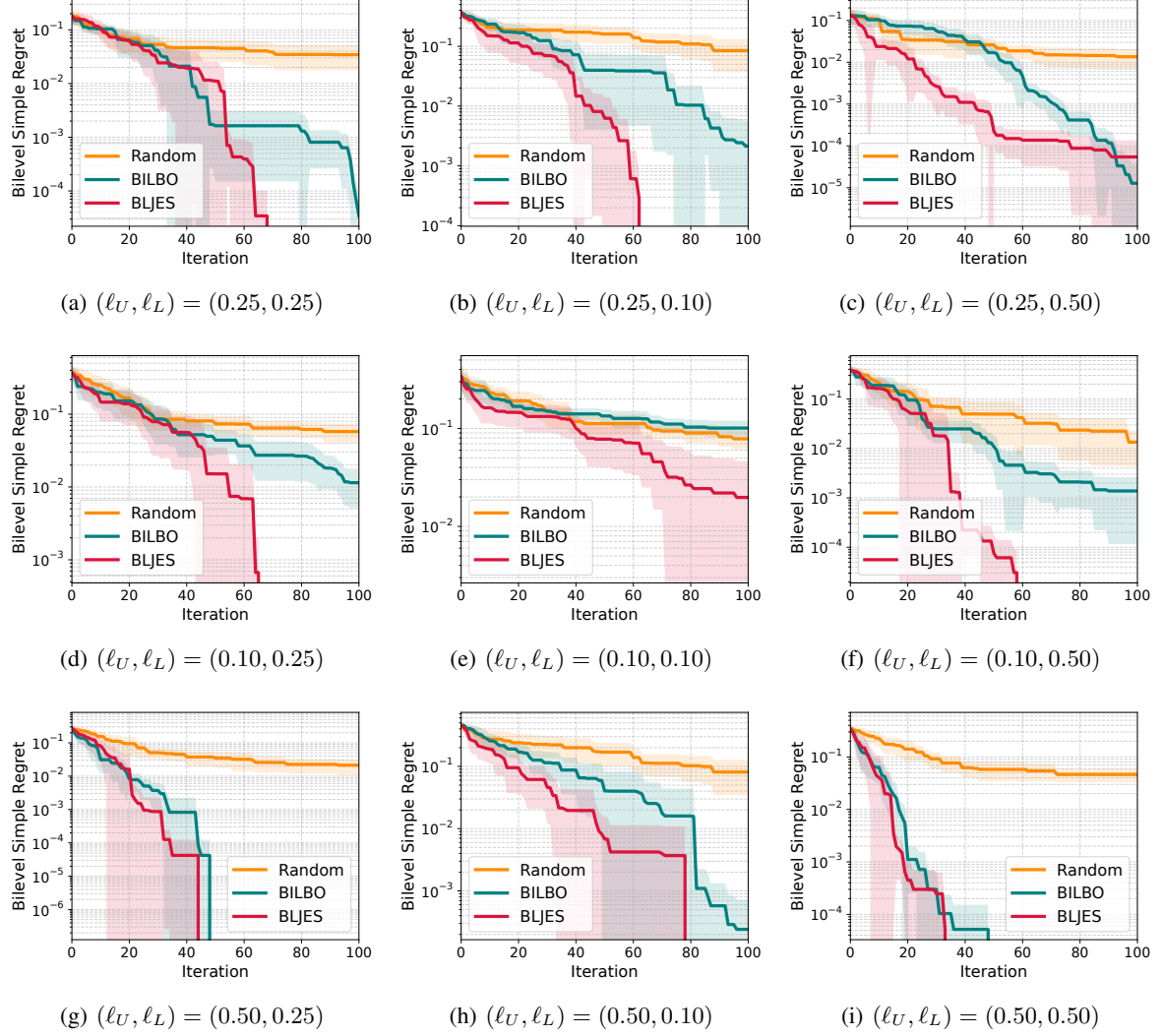


Figure 9: Regret comparison on functions from the GP prior under decoupled setting.

The results are in Fig. 13. For SMD09 and SMD11, BLJES shows faster decrease of the regret. For SMD12, BLJES and BILBO are comparable and both of them are much better than Random. For SMD10, BLJES rapidly decreased the regret, while BLJES also quickly decreased the regret (the difference is in small scale values).

E.8 HIGHER DIMENSIONAL PROBLEMS

By using the GP prior function, we evaluate performance on higher dimensional problems. Figure 14 shows the results on $d_{\mathcal{X}} = d_{\Theta} = 4$ and $d_{\mathcal{X}} = d_{\Theta} = 5$. Since creating fine grid points is difficult for these settings, the query setting (continuous domain) is employed here, because of which BILBO is not shown. We see that BLJES shows reasonable performance on the four dimension problems (a)-(d), while performance difference from the random selection becomes unclear on the five dimension problems (a)-(d).

In general, it is widely known that higher dimensional problems (e.g., more than 10 dimension) are difficult for BO [e.g., Santoni et al., 2024]. In our bilevel problem, the dimension of the search space is $d_{\mathcal{X}} + d_{\Theta}$, while the surrogate model is estimated on $d_{\mathcal{X}}$ and d_{Θ} spaces separately. We conjecture that our results are consistent with the general consensus of the high dimensional performance of BO by considering the additional difficulty caused by the low-level problem optimality constraint. Many studies exist for dealing with high dimensional problems [e.g., reviewer by Malu et al., 2021, González-Duque et al., 2024], but the efficient high dimensional exploration is still an important open problem in the context

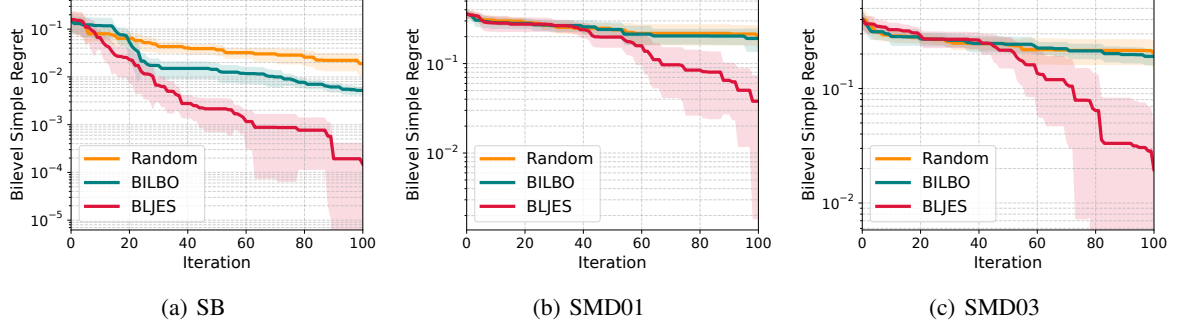


Figure 10: Regret comparison on additional benchmark problems under decoupled setting.

of BO. Some existing strategies for high dimensional problems are applicable to BLJES. For example, the well-known random projection-based method called REMBO [Wang et al., 2016] is applicable just by preparing two random projections for $\mathbf{x}$ and $\boldsymbol{\theta}$. Another well-known general strategy is LineBO [Kirschner et al., 2019] that explores one-dimensional (or low-dimensional) subspace at each iteration. In the case of BLJES, by defining one-dimensional subspace for each of $\mathbf{x}$ and $\boldsymbol{\theta}$, the same procedure as LineBO can be performed. However, detailed investigation for high dimensional setting is out of scope of this paper and it should be an important future direction.

E.9 BLJES WITHOUT TRUNCATION CONDITIONS

One of notable properties of BLJES is to consider the two truncations $f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*$ and $g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*$ as described in the first and the second items in the itemization after (4). Intuitively, these conditions transmit the information that f^* and g^* are the maximum values of the upper- and the lower- problems to the query point $(\mathbf{x}, \boldsymbol{\theta})$. To see the contributions of these truncations, we consider a simplified variant that removes the truncation conditions ($f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*$ and $g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*$) from the BLJES criterion (5), as shown in

$$\text{LB}_{\text{wot}}(\mathbf{x}, \boldsymbol{\theta}) := \mathbb{E}_{\Omega} \left[\log \frac{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f | \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f | \mathcal{D}_t)} + \log \frac{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g | \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g | \mathcal{D}_t)} \right].$$

An empirical comparison with the original BLJES is shown in Fig. 15. We can clearly see that ‘w/o truncation’ is much worse than ‘w/ truncation’. This indicates that the truncation conditions largely contributes to represent effectiveness of each candidate point.

Note that, from the viewpoint of the independence approximation, BLJES without truncations can be seen as a variational lower bound with a stronger independence assumption as described in Appendix F.2.

E.10 EMPIRICAL EVALUATION OF EFFECT OF APPROXIMATION

E.10.1 Random Fourier Approximation

We use RFF for sampling Ω in the BLJES computation as described in section 3.2. In the case of the pool setting (finite domain), the sampling from the original GPs of f and g is also possible as far as the sizes of $|\mathcal{X}|$ and $|\Theta|$ are moderate (because direct implementation of the sampling from the GP posterior requires $O(|\mathcal{X}|^3)$ and $O(|\Theta|^3)$). Therefore, based on the pool setting that is used in section 5, we here evaluate performance difference between BLJES with the original GP posterior sampling and that with the RFF-based sampling. The results on the BG benchmark are in Fig. 16. We see that the transition of the regret is highly similar, which suggests that RFF did not have large effect on the performance in this dataset.

E.10.2 Combined Error by MC approximation and Random Fourier Feature

We here evaluate the quality of computations to approximate the true lower bound (5). As described in section 3.2, we use the MC approximation for the expectation $\mathbb{E}_{\Omega}$, and RFF is also used for sampling the optimal solutions and values. The

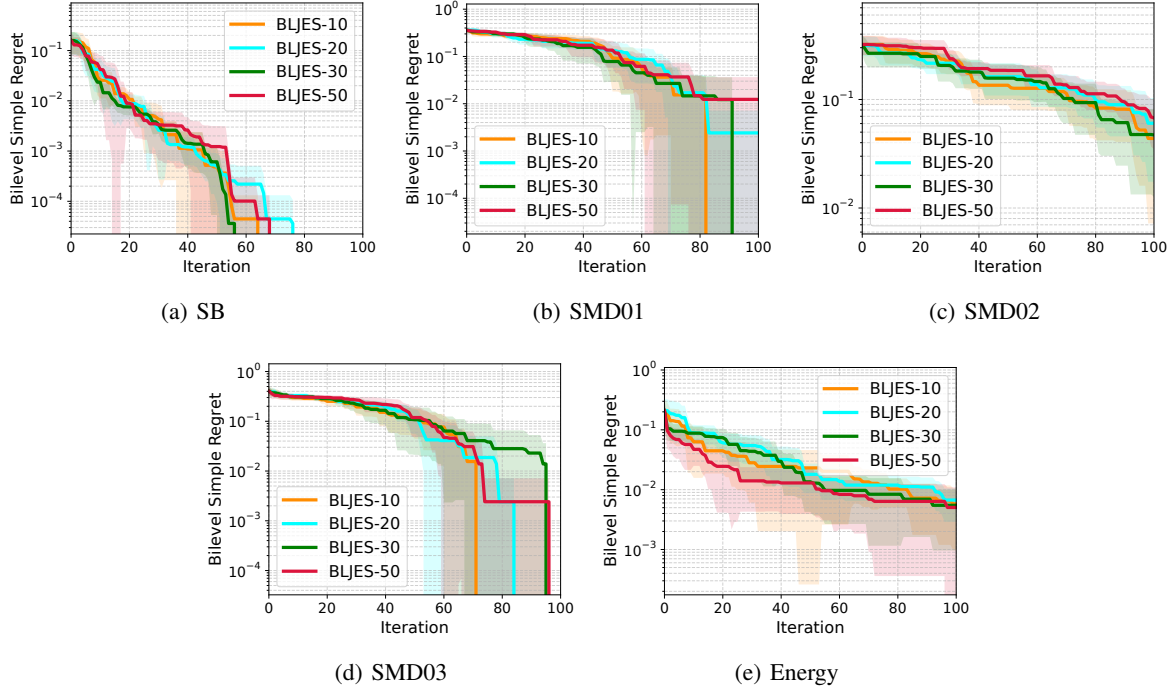


Figure 11: BLJES with different K .

pseudo ground truth of (5) was created by using the sampling from the original GP posteriors of f and g with the number of samplings $K = 10^4$. The mean squared error compared with this pseudo ground truth is shown in Fig. 17 (10 trials). ‘Exact’ indicates the GP posterior, and RFF-1000 and RFF-500 are RFF $D = 1000$ and $D = 500$, respectively. Overall, ‘Exact’ shows lower errors, while the error of RFF decreased for the larger D . When K increases, the error of RFF converges to some non-zero value, which approximately represents the error purely caused by RFF. We see that the error rapidly decreased with the increase of K , which suggests fast convergence of the MC sampling. As a results, we do not see a severe deterioration by the combination of the MC approximation and RFF.

E.11 RUN TIME EVALUATION OF ACQUISITION FUNCTIONS

In Fig. 18, we evaluated the runtime of acquisition function optimization on the BG benchmark at the initial iteration for different numbers of MC samples K . In BLJES, the computational time was almost linearly increased with K . BILBO does not rely on an MC sampling approximation and it was faster than BLJES. However, we see that computational cost of BLJES is sufficiently acceptable for practical use.

F LOWER BOUND AS INDEPENDENCE APPROXIMATION

Here, we show a connection between our variational distribution (4) and a conditional independence approximation. In a variational inference, it is common to introduce independence assumptions for making computations tractable. Our truncation based computations can also be interpreted by a similar perspective that has not been revealed in other studies of information-theoretic BO.

The variational distribution $q(y_{(x,\theta)}^f, y_{(x,\theta)}^g \mid o^*, \mathcal{D}_t)$ is a surrogate for $p(y_{(x,\theta)}^f, y_{(x,\theta)}^g \mid o^*, \mathcal{D}_t)$. In BLJES, the conditioning of o^* (i.e., f^*, g^*, x^* and θ^*) is represented by the following three conditions:

- C1. When f^* is given, $f(x', \theta^*(x')) \leq f^*$ should hold for $\forall x'$.
- C2. When g^* is given, $g(x^*, \theta') \leq g^*$ should hold for $\forall \theta'$.
- C3. $f(x^*, \theta^*) = f^*$ and $g(x^*, \theta^*) = g^*$

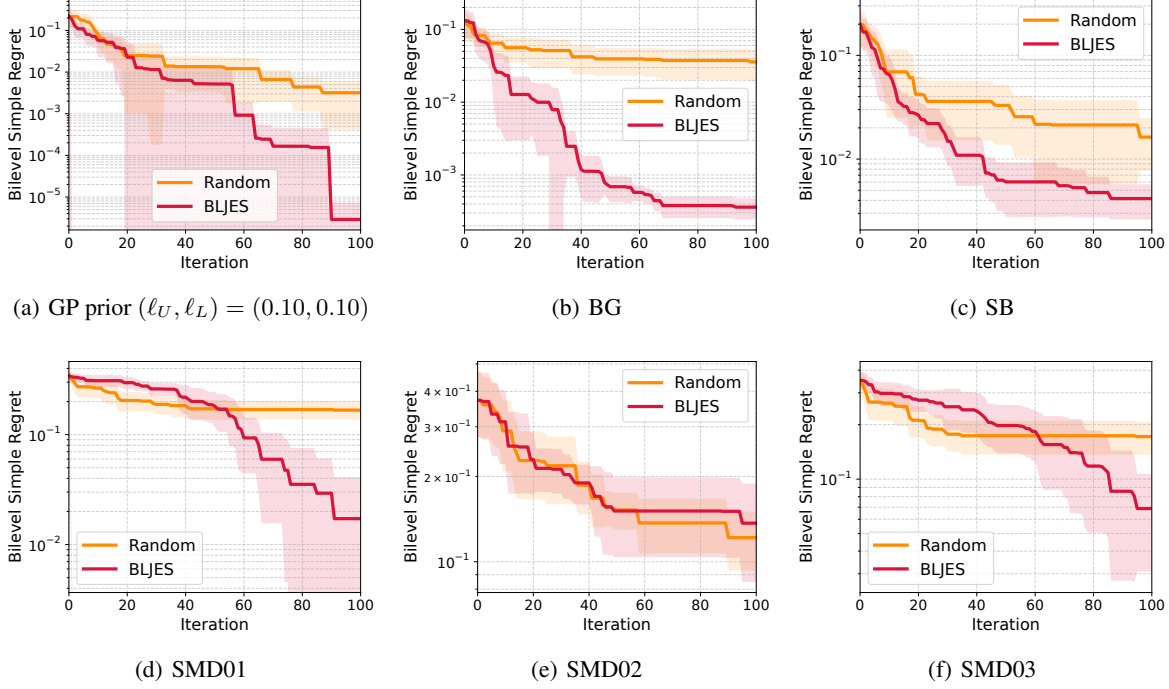


Figure 12: Regret comparison on continuous input domain.

We show that these (simplified) conditions can be connected to the independence approximation.

F.1 DERIVING BLJES THROUGH INDEPENDENCE APPROXIMATION

Let f' be f in which $(\mathbf{x}, \boldsymbol{\theta})$ and $(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))$ are removed from its input domain, and g' be g in which $(\mathbf{x}, \boldsymbol{\theta})$ and $(\mathbf{x}^*, \boldsymbol{\theta})$ are removed from its input domain. We introduce the following three independence approximations:

- Independence approximation 1 (A1): $f \perp g \mid o^*, \mathcal{D}_t$
- Independence approximation 2 (A2): $\{f(\mathbf{x}, \boldsymbol{\theta}), f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x}))\} \perp f' \mid \mathcal{D}^+$
- Independence approximation 3 (A3): $\{g(\mathbf{x}, \boldsymbol{\theta}), g(\mathbf{x}^*, \boldsymbol{\theta})\} \perp g' \mid \mathcal{D}^+$

We will derive that MI is written as follows under the above assumptions:

$$\text{MI}(y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^g; o^* \mid \mathcal{D}_t) = \mathbb{E}_\Omega \left[\log \frac{\mathbb{E}_{\boldsymbol{\theta}^*(\mathbf{x}) \mid o^*, \mathcal{D}_t} [p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*, \mathcal{D}_t^+)] p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t)} \right]$$

In BLJES, the MC approximation of the outer expectation is performed and one sample approximation is performed for the inner expectation $\mathbb{E}_{\boldsymbol{\theta}^*(\mathbf{x}) \mid o^*, \mathcal{D}_t}$ by using a sample obtained during the MC approximation of the outer expectation $\mathbb{E}_\Omega$.

The first assumption A1 indicates that upper- and lower- objective are independent given o^* and $\mathcal{D}_t$. A2 and A3 correspond to C1 and C2, which we will see the connection in the later derivation. Under the assumption of A1,

$$p(f, g \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) = p(f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) p(g \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \quad (15)$$

Note that these are the posterior distributions over the ‘entire functions’ f and g (not only for given specific input points).

We first consider the distribution of f in (15), which can be written as

$$\begin{aligned} p(f \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \\ = \mathbb{E}_{\boldsymbol{\theta}^*(\cdot) \mid o^*, \mathcal{D}_t} [p(f \mid \boldsymbol{\theta}^*(\cdot), f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)], \end{aligned}$$

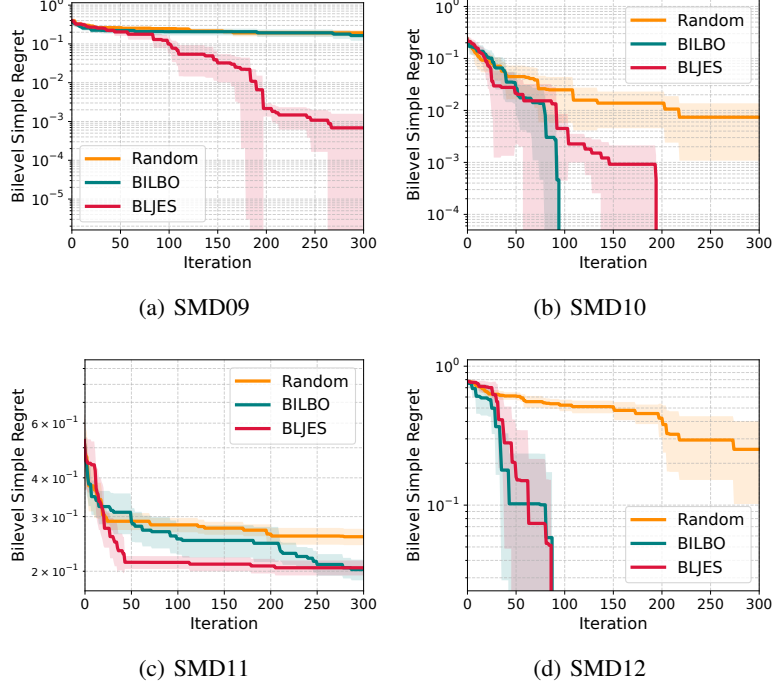


Figure 13: Regret comparison in constraint problems.

where $\theta^*(\cdot)$ is the optimal θ as a function x . The inside the expectation can be transformed into

$$\begin{aligned}
 & p(f \mid \theta^*(\cdot), f^*, g^*, x^*, \theta^*, \mathcal{D}_t) \\
 &= p(f \mid \mathcal{D}_t^+) \mathbb{I}(f(x', \theta^*(x')) \leq f^* \text{ for } \forall x' \times \text{const.}) \\
 &= p(f(x, \theta), f(x, \theta^*(x)) \mid \mathcal{D}_t^+) \mathbb{I}(f(x, \theta^*(x)) \leq f^*) p(f' \mid \mathcal{D}_t^+) \mathbb{I}(f(x', \theta^*(x')) \leq f^* \text{ for } \forall x' \neq x) \times \text{const.},
 \end{aligned}$$

where the last equality is from the assumption A2. By considering the marginal distribution, we have

$$\begin{aligned}
 & p(f(x, \theta) \mid \theta^*(\cdot), f^*, g^*, x^*, \theta^*, \mathcal{D}_t) \\
 &= \int_{f(x, \theta^*(x))} \int_{f'} p(f \mid \theta^*(\cdot), f^*, g^*, x^*, \theta^*, \mathcal{D}_t) df' df(x, \theta^*(x)) \\
 &= \int_{f(x, \theta^*(x))} p(f(x, \theta), f(x, \theta^*(x)) \mid \mathcal{D}_t^+) \mathbb{I}(f(x, \theta^*(x)) \leq f^*) df(x, \theta^*(x)) \times \\
 & \quad \underbrace{\int_{f'} p(f' \mid \mathcal{D}_t^+) \mathbb{I}(f(x', \theta^*(x')) \leq f^* \text{ for } \forall x' \neq x) df'}_A \times \text{const.} \\
 &= p(f(x, \theta) \mid f(x, \theta^*(x)) \leq f^*, \mathcal{D}_t^+),
 \end{aligned}$$

where $\int_{f'} \cdot df'$ is marginalization over the entire input domain of f' , and the last equality is from that A does not depend on $f(x, \theta)$. Therefore, we have

$$p(f \mid f^*, g^*, x^*, \theta^*, \mathcal{D}_t) = \mathbb{E}_{\theta^*(x) \mid o^*, \mathcal{D}_t} [p(f(x, \theta) \mid f(x, \theta^*(x)) \leq f^*, \mathcal{D}_t^+)]. \quad (16)$$

Next, we consider the distribution of g in (15),

$$\begin{aligned}
 & p(g \mid f^*, g^*, x^*, \theta^*, \mathcal{D}_t) \\
 &= p(g \mid \mathcal{D}_t^+) \mathbb{I}(g(x^*, \theta') \leq g^* \text{ for } \forall \theta') \times \text{const.} \\
 &= p(g(x, \theta), g(x^*, \theta) \mid \mathcal{D}_t^+) \mathbb{I}(g(x^*, \theta) \leq g^*) p(g' \mid \mathcal{D}_t^+) \mathbb{I}(g'(x^*, \theta') \leq g^* \text{ for } \forall \theta' \neq \theta) \times \text{const.},
 \end{aligned}$$

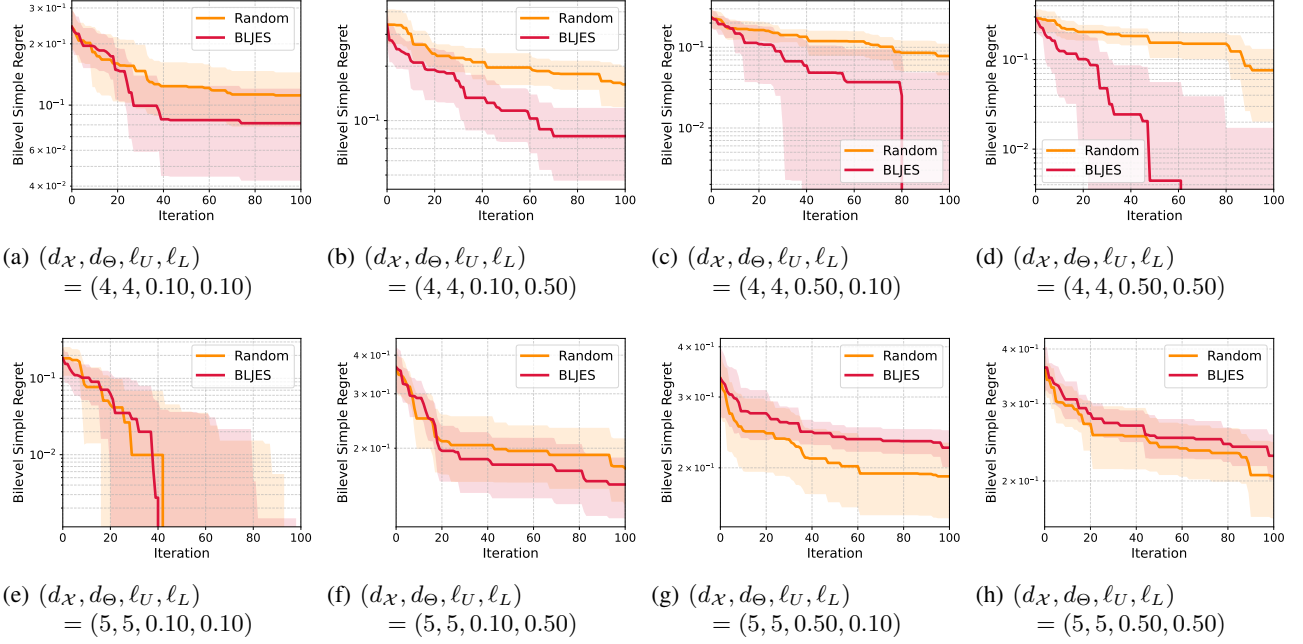


Figure 14: Regret comparison on functions from the GP prior.

where the last equality is under the assumption of A3. By considering the marginal distribution,

$$\begin{aligned}
 & p(g(\mathbf{x}, \boldsymbol{\theta}) \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \\
 &= \int_{g(\mathbf{x}^*, \boldsymbol{\theta})} \int_{g'} p(g \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) dg' dg(\mathbf{x}^*, \boldsymbol{\theta}) \\
 &= \int_{g(\mathbf{x}^*, \boldsymbol{\theta})} p(g(\mathbf{x}, \boldsymbol{\theta}), g(\mathbf{x}^*, \boldsymbol{\theta}) \mid \mathcal{D}_t^+) \mathbb{I}(g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*) dg(\mathbf{x}^*, \boldsymbol{\theta}) \times \underbrace{\int_{g'} p(g' \mid \mathcal{D}_t^+) \mathbb{I}(g'(\mathbf{x}^*, \boldsymbol{\theta}') \leq g^* \text{ for } \forall \boldsymbol{\theta}' \neq \boldsymbol{\theta}) dg'}_B \times \text{const.} \\
 &= p(g(\mathbf{x}, \boldsymbol{\theta}) \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+).
 \end{aligned} \tag{17}$$

The last equality is from that B does not depend on $g(\mathbf{x}, \boldsymbol{\theta})$.

By substituting (16) and (17) into (15),

$$p(f(\mathbf{x}, \boldsymbol{\theta}), g(\mathbf{x}, \boldsymbol{\theta}) \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) = \mathbb{E}_{\boldsymbol{\theta}^*(\mathbf{x}) \mid o^*, \mathcal{D}_t} [p(f(\mathbf{x}, \boldsymbol{\theta}) \mid f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*, \mathcal{D}_t^+)] p(g(\mathbf{x}, \boldsymbol{\theta}) \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+).$$

Then, from the definition of the observation noise,

$$\begin{aligned}
 & p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \\
 &= p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f(\mathbf{x}, \boldsymbol{\theta})) p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid g(\mathbf{x}, \boldsymbol{\theta})) p(f(\mathbf{x}, \boldsymbol{\theta}), g(\mathbf{x}, \boldsymbol{\theta}) \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \\
 &= \mathbb{E}_{\boldsymbol{\theta}^*(\mathbf{x}) \mid o^*, \mathcal{D}_t} [p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*, \mathcal{D}_t^+)] p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+).
 \end{aligned}$$

As a result, under the assumption A1, A2, and A3,

$$\text{MI}(y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^g; o^* \mid \mathcal{D}_t) = \mathbb{E}_{\Omega} \left[\log \frac{\mathbb{E}_{\boldsymbol{\theta}^*(\mathbf{x}) \mid o^*, \mathcal{D}_t} [p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid f(\mathbf{x}, \boldsymbol{\theta}^*(\mathbf{x})) \leq f^*, \mathcal{D}_t^+)] p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid g(\mathbf{x}^*, \boldsymbol{\theta}) \leq g^*, \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t)} \right]$$

As already mentioned, BLJES performs one sample approximation for the inner expectation. To perform more exact MC estimation, the nested MC approximation is required for the inner expectation, but we employ the simple one sample approximation because of the computational difficulty of the sample approximation of $\mathbb{E}_{\boldsymbol{\theta}^*(\mathbf{x}) \mid o^*, \mathcal{D}_t}$. Clarifying the error

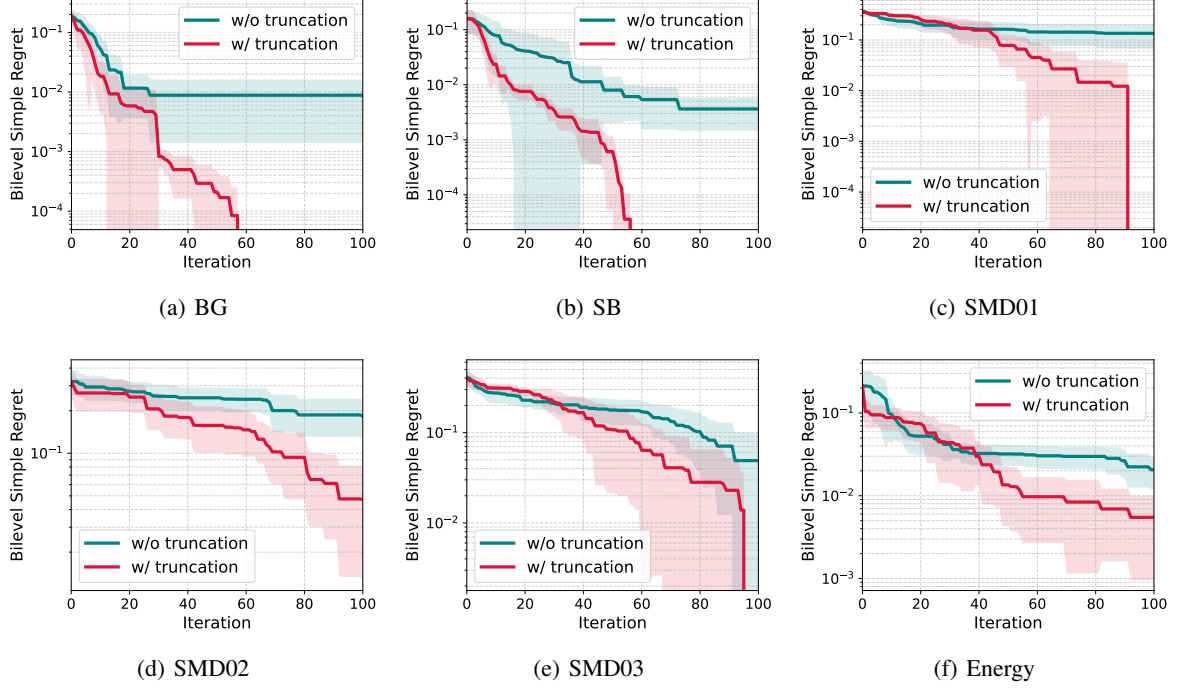


Figure 15: Evaluation of effect of truncation in BLJES on benchmark problems.

caused by this independence assumption is still future work. However, this analysis so far reveals that the ‘truncation’ based replacement of the conditioning on o^* can be connected to the independence assumption, which has never been discussed in existing information-theoretic BO studies.

F.2 DERIVING SIMPLIFIED BLJES WITHOUT TRUNCATION

By introducing stronger independence assumptions than A2 and A3 shown in the beginning of Appendix F.1, we can derive a simplified variant of BLJES, which we use for empirically evaluating the importance of truncations in Appendix E.9.

In this simplified variant, instead of A2 and A3, the conditional independence is assumed between $(\mathbf{x}, \boldsymbol{\theta})$ and all the other input points, i.e.,

- Independence approximation 2' (A2'): $f(\mathbf{x}, \boldsymbol{\theta}) \perp f' \mid \mathcal{D}^+$
- Independence approximation 3' (A3'): $g(\mathbf{x}, \boldsymbol{\theta}) \perp g' \mid \mathcal{D}^+$

where f' and g' are f and g in which only $(\mathbf{x}, \boldsymbol{\theta})$ is removed from the input domain (note that the definitions are slightly different from previous Appendix F.1). This results in

$$p(f \mid \mathcal{D}_t^+) \approx p(f(\mathbf{x}, \boldsymbol{\theta}) \mid \mathcal{D}_t^+) p(f' \mid \mathcal{D}_t^+), \quad (18)$$

$$p(g \mid \mathcal{D}_t^+) \approx p(g(\mathbf{x}, \boldsymbol{\theta}) \mid \mathcal{D}_t^+) p(g' \mid \mathcal{D}_t^+), \quad (19)$$

In this case, based on the same idea as Appendix F.1, it is easy to derive

$$p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t) \approx p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t^+) p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t^+).$$

We can define the acquisition function by defining $q(y_{(\mathbf{x}, \boldsymbol{\theta})}^f, y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid f^*, g^*, \mathbf{x}^*, \boldsymbol{\theta}^*, \mathcal{D}_t)$ as the right hand side of this approximation. As a result, we obtain a simpler lower bound (a lower bound without truncation)

$$\text{LB}_{\text{wot}}(\mathbf{x}, \boldsymbol{\theta}) := \mathbb{E}_{\Omega} \left[\log \frac{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^f \mid \mathcal{D}_t)} + \log \frac{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t^+)}{p(y_{(\mathbf{x}, \boldsymbol{\theta})}^g \mid \mathcal{D}_t)} \right], \quad (20)$$

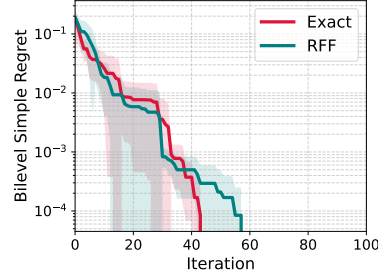


Figure 16: Evaluation of BLJES with the original GP posterior sampling and with RFF-based sampling (BG benchmark).

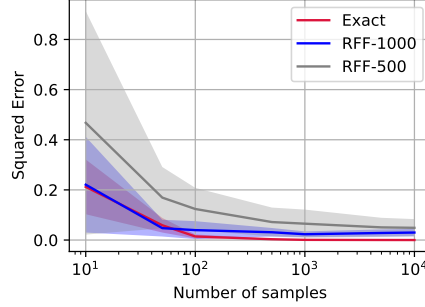


Figure 17: Approximation error of MC sampling and RFF (BG benchmark).

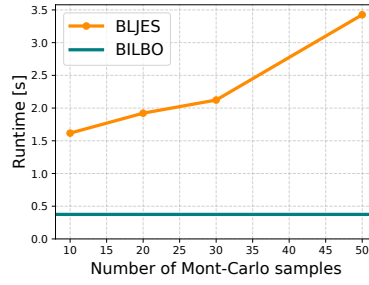


Figure 18: The CPU time (sec) of the acquisition function maximization in the BG benchmark function (10 runs average). This plot excludes the time of the GP posterior estimation because it is common for BLJES and BILBO.

for which the same Monte-Carlo approximation as the original BLJES can be applied. By comparing $LB_{\text{wot}}(\mathbf{x}, \boldsymbol{\theta})$ with (5), we see that the truncation conditions are removed.

G EXISTING STUDIES FOR REGRET ANALYSIS OF INFORMATION-THEORETIC BO

The well-known max-value entropy search (MES) [Wang and Jegelka, 2017] provides the regret bound for the special case in which only one Monte-Carlo (MC) sample is used for its expectation approximation. However, technical problems in their proof were pointed out by [Takeno et al., 2022b]. Takeno et al. [2024] provided the regret bound of an acquisition function equivalent to the one sample MES, but it is still for the one sample special case that cannot be applied to the usual MC approximation by multiple samples. In the context of information-theoretic BO for (single-level) multi-objective optimization, Belakaria et al. [2019] discussed a regret bound, but Suzuki et al. [2020] pointed out an obvious significant mistake that can make the regret a negative value.