Multi-fidelity Bayesian Optimization with Max-value Entropy Search and its Parallelization

Shion Takeno [1 2] **Hitoshi Fukuoka** [3] **Yuhki Tsukada** [3 4] **Toshiyuki Koyama** [3] **Motoki Shiga** [2 4 5]
Ichiro Takeuchi [1 2] **Masayuki Karasuyama** [1 4 6]

Abstract

In a standard setting of Bayesian optimization (BO), the objective function evaluation is assumed to be highly expensive. Multi-fidelity Bayesian optimization (MFBO) accelerates BO by incorporating lower fidelity observations available with a lower sampling cost. We propose a novel information-theoretic approach to MFBO, called multi-fidelity max-value entropy search (MF-MES), that enables us to obtain a more reliable evaluation of the information gain compared with existing information-based methods for MFBO. Further, we also propose a parallelization of MF-MES mainly for the asynchronous setting because queries typically occur asynchronously in MFBO due to a variety of sampling costs. We show that most of computations in our acquisition functions can be derived analytically, except for at most only two dimensional numerical integration that can be performed efficiently by simple approximations. We demonstrate effectiveness of our approach by using benchmark datasets and a real-world application to materials science data.

1. Introduction

Bayesian optimization (BO) is a popular machine-learning technique for the black-box optimization problem. Efficiency of BO has been widely shown in a variety of application areas such as scientific experiments (Wigley et al., 2016), simulation calculations (Ramprasad et al., 2017), and tuning of machine-learning methods (Snoek et al., 2012). In these scenarios, observing an objective function value is usually quite expensive and thus achieving the optimal value with low querying cost is strongly demanded.

Although standard BO only considers directly querying to an objective function $f(\boldsymbol{x})$, in many practical problems, lower fidelity approximations of the original objective function can be observed. For example, theoretical computations of physical processes often have multiple levels of approximations by which the trade-off between the computational cost and accuracy can be controlled. The goal of *multi-fidelity Bayesian optimization* (MFBO) is to accelerate BO by utilizing those lower fidelity observations to reduce the total cost of the optimization.

In this paper, we focus on the information-based approach. For usual BO without multi-fidelity, which we call *single fidelity BO*, seminal works of this direction are *entropy search* (ES) and *predictive entropy search* (PES) proposed by Hennig & Schuler (2012) and Hernández-Lobato et al. (2014), respectively. They define acquisition functions by using information gain for the optimal solution $\boldsymbol{x}_* := \mathrm{argmax}_{\boldsymbol{x}} f(\boldsymbol{x})$. Unlike classical evaluation measures such as expected improvement, the information-based criterion is a measure of global utility which does not require any additional exploit-explore trade-off parameter. The superior performance of information-based methods have been shown empirically, and then, the same approach has also been extended to the multi-fidelity setting (Swersky et al., 2013; Zhang et al., 2017).

Even in the case of single fidelity BO, however, accurately evaluating information gain is notoriously difficult, which often requires complicated numerical approximations. For MFBO, evaluating information across multiple fidelities is further difficult. To overcome this difficulty, we consider a novel information-based approach to MFBO, which is based on a variant of ES called *max-value entropy search* (MES), proposed by Wang & Jegelka (2017). MES considers the information gain for $f_* := \max_{\boldsymbol{x}} f(\boldsymbol{x})$ instead

[1] Department of Computer Science, Nagoya Institute of Technology, Aichi, Japan [2] Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan [3] Department of Materials Design Innovation Engineering, Nagoya University, Aichi, Japan [4] PRESTO, Japan Science and Technology Agency, Saitama, Japan [5] Department of Electrical, Electronic and Computer Engineering, Gifu University, Gifu, Japan [6] Center for Materials Research by Information Integration, National Institute for Material Science, Ibaraki, Japan. Correspondence to: M. Karasuyama <karasuyama@nitech.ac.jp>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

of $\boldsymbol{x}_*$. This greatly facilitates the computation of the information gain because f_* is in one dimensional space unlike $\boldsymbol{x}_*$, and they showed superior performance of MES compared with ES/PES. Our method, called *multi-fidelity MES* (MF-MES), can evaluate the information gain for f_* from an observation of an arbitrary fidelity, and we show that additional expressions, compared with MES, can be derived analytically except for one dimensional integral, which can be calculated accurately and efficiently by using standard numerical integration techniques. This enables us to obtain a more reliable evaluation of the information gain compared with existing information-based MFBO methods containing approximations that are difficult to justify. Our MF-MES is also advantageous to other measures of global utility for MFBO, such as the knowledge gradient-based method (Poloczek et al., 2017), because they are often computationally extremely complicated. Section 5 discusses related studies in more detail.

Further, we also propose *parallelization* of MF-MES. Since objective functions have a variety of sampling costs, queries naturally occur *asynchronously* in MFBO. We extend our information gain so that points currently being queried can be taken into consideration. Similarly in the case of MF-MES, we show that a required numerical integration in addition to the sampling of f_* is also reduced to one dimensional space through the integration by substitution. This allows us to obtain the reliable evaluation of the information gain for the parallel extension of MF-MES.

Our main contributions are summarized as follows:

1. We develop an information-theoretic efficient MFBO method. Naïve formulation and implementation of this problem raise computationally challenging issues that need to be addressed by carefully-tuned and time-consuming approximate computations. By using several computational tricks mainly inspired by MES (Wang & Jegelka, 2017), we show that this computational bottleneck can be nicely avoided without additional assumptions or approximations.

2. We develop an information-theoretic asynchronous parallel MFBO method. To our knowledge, there are no existing works in this topic — We believe that our method is useful in many practical experimental design and black-box optimization tasks with multiple information sources with different fidelities and its parallel evaluation.

We empirically demonstrate effectiveness of our approach by using benchmark functions and a real-world application to materials science data.

2. Preliminary

In this section, we first briefly review a multi-fidelity extension of *Gaussian process regression* (GPR). Suppose that $y_{\boldsymbol{x}}^{(1)}, \ldots, y_{\boldsymbol{x}}^{(M)}$ are the observations at $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ with M different fidelities in which $y_{\boldsymbol{x}}^{(M)}$ is the highest fidelity and $y_{\boldsymbol{x}}^{(1)}$ is the lowest fidelity. Each observation is modeled as $y_{\boldsymbol{x}}^{(m)} = f_{\boldsymbol{x}}^{(m)} + \epsilon$ in which a random noise $\epsilon \sim \mathcal{N}(0, \sigma_{\text{noise}}^2)$ is added to the underlying true function $f_{\boldsymbol{x}}^{(m)} : \mathcal{X} \to \mathbb{R}$. The training data set $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_{\boldsymbol{x}_i}^{(m_i)}, m_i)\}_{i \in [n]}$ contains a set of triplets consisting of an input $\boldsymbol{x}_i$, fidelity $m_i \in [M]$, and an output $y_{\boldsymbol{x}_i}^{(m_i)}$, where $[n] := \{1, \ldots, n\}$.

Throughout the paper, we assume that a set of outputs $\{f_{\boldsymbol{x}}^{(m)}\}$ for any set of pairs $(\boldsymbol{x}, m)$ are always modeled as the multi-variate normal distribution. Standard multi-output extensions of GPR such as multi-task GPR (Bonilla et al., 2008), co-kriging (Kennedy & O'Hagan, 2000), and semiparametric latent factor model (SLFM) (Teh et al., 2005), satisfy this condition. We call GPR fitted to observations across multiple fidelities *multi-fidelity Gaussian process regression* (MF-GPR), in general.

MF-GPR defines a kernel function $k((\boldsymbol{x}_i, m_i), (\boldsymbol{x}_j, m_j))$ for a pair of training instances $(\boldsymbol{x}_i, y_{\boldsymbol{x}_i}^{(m_i)}, m_i)$ and $(\boldsymbol{x}_j, y_{\boldsymbol{x}_j}^{(m_j)}, m_j)$. An example of this kernel function in the case of SLFM is shown in appendix A.1. By defining a kernel matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ in which the i, j element is defined by $k((\boldsymbol{x}_i, m_i), (\boldsymbol{x}_j, m_j))$, all the fidelities $f^{(1)}, \ldots, f^{(M)}$ are integrated into a GPR model in which predictive mean and variance are $\mu_{\boldsymbol{x}}^{(m)} = \boldsymbol{k}_n^{(m)}(\boldsymbol{x})^\top \boldsymbol{C}^{-1} \boldsymbol{y}$, and $\sigma_{\boldsymbol{x}}^{2(m)} = k((\boldsymbol{x}, m), (\boldsymbol{x}, m)) - \boldsymbol{k}_n^{(m)}(\boldsymbol{x})^\top \boldsymbol{C}^{-1} \boldsymbol{k}_n^{(m)}(\boldsymbol{x})$, where $\boldsymbol{C} := \boldsymbol{K} + \sigma_{\text{noise}}^2 \boldsymbol{I}$ with the identity matrix $\boldsymbol{I}$, $\boldsymbol{y} := (y_{\boldsymbol{x}_1}^{(m_1)}, \ldots, y_{\boldsymbol{x}_n}^{(m_n)})^\top$, and $\boldsymbol{k}_n^{(m)}(\boldsymbol{x}) := (k((\boldsymbol{x}, m), (\boldsymbol{x}_1, m_1)), \ldots, k((\boldsymbol{x}, m), (\boldsymbol{x}_n, m_n)))^\top$. For later use, we define $\sigma_{\boldsymbol{x}}^{2(mm')}$ as the predictive covariance between $(\boldsymbol{x}, m)$ and $(\boldsymbol{x}, m')$, i.e., covariance for the identical $\boldsymbol{x}$ at different fidelities: $\sigma_{\boldsymbol{x}}^{2(mm')} = k((\boldsymbol{x}, m), (\boldsymbol{x}, m')) - \boldsymbol{k}_n^{(m)}(\boldsymbol{x})^\top \boldsymbol{C}^{-1} \boldsymbol{k}_n^{(m')}(\boldsymbol{x})$.

3. Multi-fidelity Bayesian Optimization with Max-value Entropy

We consider Bayesian optimization (BO) for maximizing the highest fidelity function $f_{\boldsymbol{x}}^{(M)}$ when M different fidelities $y_{\boldsymbol{x}}^{(m)}$ for $m = 1, \ldots, M$ are available to querying. The querying cost is assumed to be known as $\lambda^{(m)}$, where $\lambda^{(1)} \leq \lambda^{(2)} \ldots \leq \lambda^{(M)}$. Our goal is to achieve a higher value with smaller accumulated cost of the queryings. We call this problem *multi-fidelity Bayesian optimization* (MFBO). When $M = 1$, MFBO is reduced to the usual black box optimization to which we refer as the single fidelity setting, while we refer to the setting $M \geq 2$ as the multi-fidelity setting.

We employ the information-based approach, which has been widely used in the single fidelity BO. In particular, our approach is inspired by *max-value entropy search* (MES) proposed by Wang & Jegelka (2017), which considers *information gain* about the optimal value $\max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ obtained by a querying. In the case of MFBO, we need to consider the information gain for identifying the maximum of the highest fidelity function $f_* := \max_{\boldsymbol{x}\in\mathcal{X}} f_{\boldsymbol{x}}^{(M)}$ by observing an arbitrary fidelity observation. We refer to our information-based MFBO as *multi-fidelity MES* (MF-MES). Although information-based approaches often result in complicated computations, we show that the calculation of our information gain is reduced to simple computations by which stable information evaluation becomes possible.

3.1. Information Gain for Sequential Querying

We first consider the case that a query is sequentially issued after the previous one is observed, which we refer to as *sequential querying*. Suppose that we already have a training data set $\mathcal{D}_t$ and need to determine next $\boldsymbol{x}_{t+1}$ and m_{t+1}. We define an acquisition function

$$a(\boldsymbol{x}, m) := I(f_*; f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t) \;/\; \lambda^{(m)}, \tag{1}$$

where $I(f_*; f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t)$ is the mutual information between f_* and $f_{\boldsymbol{x}}^{(m)}$ conditioned on $\mathcal{D}_t$. By maximizing $a(\boldsymbol{x}, m)$, we obtain a pair of the input $\boldsymbol{x}$ and the fidelity m which maximally gains information of the optimal value f_* of the highest fidelity per unit cost.

The mutual information can be written as the difference of the entropy:

$$\begin{aligned} &I(f_*; f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t) \\ &= H(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t) - \mathbb{E}_{f_* \mid \mathcal{D}_t}\big[H(f_{\boldsymbol{x}}^{(m)} \mid f_*, \mathcal{D}_t)\big], \end{aligned} \tag{2}$$

where $H(\cdot \mid \cdot)$ is the conditional entropy of $p(\cdot \mid \cdot)$. The first term in the right hand side can be derived analytically for any fidelity m: $H(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t) = \log\Big(\sigma_{\boldsymbol{x}}^{(m)}\sqrt{2\pi e}\Big)$, where $e := \exp(1)$. The second term in (2) takes the expectation over the maximum f_*. Since an analytical formula is not known for this expectation, we employ Monte Carlo estimation by sampling f_* from the current GPR:

$$\mathbb{E}_{f_* \mid \mathcal{D}_t}\big[H(f_{\boldsymbol{x}}^{(m)} \mid f_*, \mathcal{D}_t)\big] \approx \sum_{f_* \in \mathcal{F}_*} \frac{H(f_{\boldsymbol{x}}^{(m)} \mid f_*, \mathcal{D}_t)}{|\mathcal{F}_*|}, \tag{3}$$

where $\mathcal{F}_*$ is a set of sampled f_*. Note that since this sampling approximation is in one dimensional space, accurate approximation can be expected with a small amount of samples. In Section 4, we discuss computational procedures of this sampling. For a given sampled f_*, the entropy of $p(f_{\boldsymbol{x}}^{(m)} \mid f_*, \mathcal{D}_t)$ is needed to calculate in (3). To make the computation tractable, we replace this conditional distribution with $p(f_{\boldsymbol{x}}^{(m)} \mid f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t)$, i.e., conditioning only on the given $\boldsymbol{x}$ rather than requiring $f_{\boldsymbol{x}}^{(M)} \leq f_*$ for $\forall \boldsymbol{x} \in \mathcal{X}$. Note that this simplification has been employed by most of entropy-based BO methods (e.g., Hernández-Lobato et al., 2014; Wang & Jegelka, 2017) including MES, and superior performance compared with other approaches has been shown.

For any $\zeta \in \mathbb{R}$, define $\gamma_{\zeta}^{(m)}(\boldsymbol{x}) := (\zeta - \mu_{\boldsymbol{x}}^{(m)})/\sigma_{\boldsymbol{x}}^{(m)}$ as a function for scaling. When $m = M$, the density function $p(f_{\boldsymbol{x}}^{(m)} \mid f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t)$ is *truncated normal distribution*. The entropy of truncated normal distribution can be represented as (Michalowicz, 2014)

$$\begin{aligned} H(f_{\boldsymbol{x}}^{(M)} \mid f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t) = &\log\Big(\sqrt{2\pi e}\,\sigma_{\boldsymbol{x}}^{(M)}\Phi\big(\gamma_{f_*}^{(M)}(\boldsymbol{x})\big)\Big) \\ &- \frac{\gamma_{f_*}^{(M)}(\boldsymbol{x})\phi\big(\gamma_{f_*}^{(M)}(\boldsymbol{x})\big)}{2\Phi\big(\gamma_{f_*}^{(M)}(\boldsymbol{x})\big)}, \end{aligned} \tag{4}$$

where ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution.

Next, we consider the case of $m \neq M$. Unlike the case of $m = M$, the density $p(f_{\boldsymbol{x}}^{(m)} \mid f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t)$ is not the truncated normal. Since MF-GPR represents all fidelities as one unified GPR, the joint marginal distribution $p(f_{\boldsymbol{x}}^{(M)}, f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t)$ can be immediately obtained from the two dimensional predictive distribution, from which we obtain $p(f_{\boldsymbol{x}}^{(M)} \mid f_{\boldsymbol{x}}^{(m)}, \mathcal{D}_t)$ as

$$f_{\boldsymbol{x}}^{(M)} \mid f_{\boldsymbol{x}}^{(m)}, \mathcal{D}_t \sim \mathcal{N}(u(\boldsymbol{x}), s^2(\boldsymbol{x})), \tag{5}$$

where $u(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^{2(mM)}\big(f_{\boldsymbol{x}}^{(m)} - \mu_{\boldsymbol{x}}^{(m)}\big)/\sigma_{\boldsymbol{x}}^{2(m)} + \mu_{\boldsymbol{x}}^{(M)}$, and $s^2(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^{2(M)} - \big(\sigma_{\boldsymbol{x}}^{2(mM)}\big)^2/\sigma_{\boldsymbol{x}}^{2(m)}$. By using this conditional distribution, the entropy of $p(f_{\boldsymbol{x}}^{(m)} \mid f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t)$ can be written as follows:

Lemma 3.1. *Let* $Z := 1/\sigma_{\boldsymbol{x}}^{(m)}\Phi(\gamma_{f_*}^{(M)}(\boldsymbol{x}))$ *and* $\Psi(f_{\boldsymbol{x}}^{(m)}) := \Phi\big((f_* - u(\boldsymbol{x}))/s(\boldsymbol{x})\big)\phi\big(\gamma_{f_{\boldsymbol{x}}^{(m)}}^{(m)}(\boldsymbol{x})\big)$. *Then, for a given* f_*, *we obtain*

$$\begin{aligned} &H(f_{\boldsymbol{x}}^{(m)} \mid f_{\boldsymbol{x}}^{(M)} \leq f_*, \mathcal{D}_t) \\ &= -\int Z\Psi(f_{\boldsymbol{x}}^{(m)}) \log\Big(Z\Psi(f_{\boldsymbol{x}}^{(m)})\Big)\, \mathrm{d}f_{\boldsymbol{x}}^{(m)}. \end{aligned} \tag{6}$$

See Appendix B for the proof.

Lemma 3.1 indicates that the entropy is represented through the one dimensional integral over $f_{\boldsymbol{x}}^{(m)}$. Since the

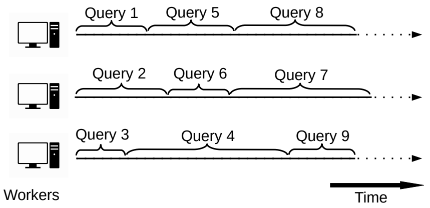


Figure 1: Asynchronous parallelization in MFBO. Because of diversity of the evaluation cost of objective functions, queries typically occur asynchronously. When a worker becomes available, a next query should be determined while taking queries being evaluated in the other workers into consideration.

integral is only on the one dimensional space, standard numerical integration techniques (e.g., quadrature) can provide precise approximation efficiently. Consequently, we see that that the entropy $H(f_{\boldsymbol{x}}^{(m)} \mid f_*, \mathcal{D}_t)$ in (3) can be obtained accurately with simple computations.

3.2. Asynchronous Parallelization

We consider an extension of MF-MES for the case that multiple queries can be issued in parallel, which we refer to as *parallel querying*. Suppose that we have $q > 1$ "workers" each one of which can evaluate an objective function value. In the context of parallel BO, the two settings called *synchronous* and *asynchronous* parallelizations can be considered. As shown in Figure 1, since MFBO evaluates a variety of different costs of objective functions, queries naturally occur asynchronously. Thus, we focus on asynchronous parallelization (See Appendix D.4 for the discussion of the synchronous setting).

Suppose that $q-1$ pairs of the input $\boldsymbol{x}$ and the fidelity m, written as $\mathcal{Q} := \{(\boldsymbol{x}_1, m_1), \ldots, (\boldsymbol{x}_{q-1}, m_{q-1})\}$, are now being evaluated by using $q-1$ workers, and an additional query to an available worker needs to be determined. Let $\boldsymbol{f}_{\mathcal{Q}} := (f_{\boldsymbol{x}_1}^{(m_1)}, \ldots, f_{\boldsymbol{x}_{q-1}}^{(m_{q-1})})^\top$. Then, a natural extension of MF-MES to determine the q-th pair $(\boldsymbol{x}_q, m_q)$ is

$$a_{\mathrm{para}}(\boldsymbol{x}, m) = I(f_*; f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}) \;/\; \lambda^{(m)}. \tag{7}$$

The numerator is the mutual information conditioned on $\boldsymbol{f}_{\mathcal{Q}}$ which is defined by

$$\begin{aligned} I(f_*; f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}) &:= \mathbb{E}_{\boldsymbol{f}_{\mathcal{Q}} \mid \mathcal{D}_t}\left[H(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}})\right] \\ &- \mathbb{E}_{\boldsymbol{f}_{\mathcal{Q}}, f_* \mid \mathcal{D}_t}\left[H(f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}, f_{\boldsymbol{x}}^{(M)} \leq f_*)\right]. \end{aligned} \tag{8}$$

Compared with the mutual information in sequential querying (2), this equation additionally takes the expectation over $\boldsymbol{f}_{\mathcal{Q}}$ which is currently under evaluation. Thus, by using (7), we can select a cost effective pair of $\boldsymbol{x}$ and m while the $q-1$ pairs running on the other workers are taken into consideration.

Although (8) contains the $|\mathcal{Q}| + 2$ dimensional integral at a glance, we show that this can be calculated by at most 2 dimensional numerical integral. Let $\boldsymbol{\Sigma}_{\mathcal{M}} \in \mathbb{R}^{2 \times 2}$ and $\boldsymbol{\Sigma}_{\mathcal{Q}} \in \mathbb{R}^{q-1 \times q-1}$ be the predictive covariance matrices for $\mathcal{M} := \{(\boldsymbol{x}, m), (\boldsymbol{x}, M)\}$ and $\mathcal{Q}$, respectively, and $\boldsymbol{\Sigma}_{\mathcal{Q},\mathcal{M}} (= \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{Q}}^\top) \in \mathbb{R}^{q-1 \times 2}$ be the predictive covariance matrix of the rows $\mathcal{Q}$ and the columns $\mathcal{M}$. For later use, we define the conditional distribution $p(f_{\boldsymbol{x}}^{(m)}, f_{\boldsymbol{x}}^{(M)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}})$ as follows

$$\begin{bmatrix} f_{\boldsymbol{x}}^{(m)} \\ f_{\boldsymbol{x}}^{(M)} \end{bmatrix} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)} \\ \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)} \end{bmatrix}, \begin{bmatrix} \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(m)} & \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)} \\ \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)} & \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(M)} \end{bmatrix} \right),$$

where

$$\begin{bmatrix} \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)} \\ \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)} \end{bmatrix} = \begin{bmatrix} \mu_{\boldsymbol{x}}^{(m)} \\ \mu_{\boldsymbol{x}}^{(M)} \end{bmatrix} + \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{Q}} \boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} (\boldsymbol{f}_{\mathcal{Q}} - \boldsymbol{\mu}_{\mathcal{Q}}), \tag{9}$$

$$\begin{bmatrix} \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(m)} & \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)} \\ \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)} & \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(M)} \end{bmatrix} = \boldsymbol{\Sigma}_{\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{Q}} \boldsymbol{\Sigma}_{\mathcal{Q}}^{-1} \boldsymbol{\Sigma}_{\mathcal{Q},\mathcal{M}}, \tag{10}$$

and $\boldsymbol{\mu}_{\mathcal{Q}} := (\mu_{\boldsymbol{x}_1}^{(m_1)}, \ldots, \mu_{\boldsymbol{x}_{q-1}}^{(m_{q-1})})^\top$. Note that (9) is a random variable vector because it depends on $\boldsymbol{f}_{\mathcal{Q}}$, while all the elements of (10) are constants. By using these equations, the mutual information (8) is re-written as follows:

Lemma 3.2. *Let*

$$\tilde{f}_* := f_* - \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}, \tag{11}$$

and $\tilde{f}_{\boldsymbol{x}}^{(m)} := f_{\boldsymbol{x}}^{(m)} - \mu_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)}$. *Then, we obtain*

$$\begin{aligned} I(f_*; f_{\boldsymbol{x}}^{(m)} \mid \mathcal{D}_t, \boldsymbol{f}_{\mathcal{Q}}) &= \log\left(\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)} \sqrt{2\pi e}\right) \\ &- \mathbb{E}_{\tilde{f}_* \mid \mathcal{D}_t}\left[\int -\eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \log \eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \, \mathrm{d}\tilde{f}_{\boldsymbol{x}}^{(m)}\right] \end{aligned} \tag{12}$$

where

$$\eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) := \frac{\Phi\left(\dfrac{\tilde{f}_* - \left(\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)} \;/\; \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(m)}\right) \tilde{f}_{\boldsymbol{x}}^{(m)}}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(M)} - \left(\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(mM)}\right)^2 \;/\; \sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{2(m)}}\right) \phi\left(\dfrac{\tilde{f}_{\boldsymbol{x}}^{(m)}}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)}}\right)}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{m} \Phi\left(\dfrac{\tilde{f}_*}{\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(M)}}\right)}. \tag{13}$$

See Appendix D.1 for the proof. It should be noted that the second term of (12) only contains the integral over two variables ($\tilde{f}_{\boldsymbol{x}}^{(m)}$ and $\tilde{f}_*$) unlike the original formulation (8). The first term of (12) can be directly calculated because $\sigma_{\boldsymbol{x}|\boldsymbol{f}_{\mathcal{Q}}}^{(m)}$ does not depend on the random vector $\boldsymbol{f}_{\mathcal{Q}}$ as shown in (10). We calculate the expectation in the second term of (12) by using the Monte Carlo estimation with sampled $\tilde{f}_*$:

$$\sum_{\tilde{f}_* \in \tilde{\mathcal{F}}_*} \frac{1}{|\tilde{\mathcal{F}}_*|} \int -\eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \log \eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)}) \, \mathrm{d}\tilde{f}_{\boldsymbol{x}}^{(m)} \tag{14}$$

where $\widetilde{\mathcal{F}}_*$ is a set of sampled $\tilde{f}_*$. The integral in this equation can be easily evaluated by using quadrature because it is on the one dimensional space and $\eta(\tilde{f}_*, \tilde{f}_{\boldsymbol{x}}^{(m)})$ can be analytically calculated from the definition (13). Further, when $m = M$, this integral is also can be analytically calculated (See Appendix D.2).

4. Computations

Algorithm 1 shows the procedure of MF-MES for sequential querying. As the first step in the every iteration, a set of max values $\mathcal{F}_*$ are sampled from $p(f_* \mid \mathcal{D}_t)$. There are several approaches to sampling the max value. Wang & Jegelka (2017) showed that the effective approximation is possible by using sampling through Gumbel distribution or random feature map (RFM). Gumbel distribution is widely known in extreme value theory (Gumbel, 1958) as one of generalized extreme value distributions.

Although the Gumbel approximation is performed under an independent approximation of GPR, Wang & Jegelka (2017) showed the accurate approximation can be obtained. In contrast, RFM (Rahimi & Recht, 2008) can incorporate dependency in the GPR model by using a set of predefined basis functions $\boldsymbol{\phi}(\boldsymbol{x}, m) \in \mathbb{R}^D$, and the highest fidelity function is represented as $f_{\boldsymbol{x}}^{(M)} \approx \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, M)$, where $\boldsymbol{w} \in \mathbb{R}^D$ (Appendix A.2 shows an example of an RFM approximation in the case of SLFM). The max value is sampled by maximizing $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, M)$ with respect to $\boldsymbol{x}$. For further detail of these two approaches, see (Wang & Jegelka, 2017), in which it is also shown that MES is empirically robust with respect to this sampling, and theoretically, they showed that the regret bound can be guaranteed even only for one sample of f_*.

Once $\mathcal{F}_*$ is generated, the acquisition function calculation can be analytically performed except for one dimensional numerical integration. Although most complicated process in the algorithm is the calculation of (6) shown in line 15 of Algoirthm 1, this is also quite simple in practice as described below. For a given f_* and the conditional distribution (5) which is constructed from the two dimension GPR predictive distribution $p(f_{\boldsymbol{x}}^{(M)}, f_{\boldsymbol{x}}^{(m)} \mid \boldsymbol{x}, \mathcal{D}_t)$, the integral of (6) can be computed by $O(1)$. Further, since (5) does not depend on sampled f_*, it is not required to re-calculate (5) for each one of sampled f_*.

For the acquisition function maximization (argmax in line 4), if the candidate space $\mathcal{X}$ is a discrete set, we simply calculate the acquisition values for all $\boldsymbol{x} \in \mathcal{X}$. For a continuous space, popular approaches such as DIRECT (Jones et al., 1993) and gradient-based optimizers are applicable. Note that our acquisition function is differentiable, and the derivative of the integral (6) can be calculated by the same one dimensional numerical integral procedure.

Algorithm 1 MF-MES for sequential querying

```
 1: function MF-MES($\mathcal{D}_0, M, \mathcal{X}, \{\lambda^{(m)}\}_{m=1}^M$)
 2:     for $t = 0, \dots, T$ do
 3:         Generate $\mathcal{F}_*$ from current $f^{(M)}(\boldsymbol{x})$
 4:         $(\boldsymbol{x}_{t+1}, m_{t+1}) \leftarrow$ argmax$_{\boldsymbol{x}\in\mathcal{X},m}$
              InfoGain($\boldsymbol{x}, m, \mathcal{F}_*, \mathcal{D}_t$) / $\lambda^{(m)}$
 5:         $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup (\boldsymbol{x}_{t+1}, y^{(m_{t+1})}(\boldsymbol{x}_{t+1}), m_{t+1})$
 6:     end for
 7: end function
 8: function InfoGain($\boldsymbol{x}, m, \mathcal{F}_*, \mathcal{D}_t$)
 9:     Calculate $\mu_{\boldsymbol{x}}^{(m)}$ and $\sigma_{\boldsymbol{x}}^{(m)}$
10:     Set $H_0 \leftarrow \log\left(\sigma_{\boldsymbol{x}}^{(m)}\sqrt{2\pi e}\right)$
11:     if $m = M$ then
12:         Set $H_1 \leftarrow \sum_{f_*\in\mathcal{F}_*} \frac{H(f_{\boldsymbol{x}}^{(M)} | f_{\boldsymbol{x}}^{(M)} \le f_*, \mathcal{D}_t)}{|\mathcal{F}_*|}$
              by using (4)
13:     else
14:         Calculate $\mu_{\boldsymbol{x}}^{(M)}$ and $\sigma_{\boldsymbol{x}}^{(M)}$ and $\sigma_{\boldsymbol{x}}^{2(mM)}$
15:         Set $H_1 \leftarrow \sum_{f_*\in\mathcal{F}_*} \frac{H(f_{\boldsymbol{x}}^{(m)} | f_{\boldsymbol{x}}^{(M)} \le f_*, \mathcal{D}_t)}{|\mathcal{F}_*|}$
              by using (6)
16:     end if
17:     Return $H_0 - H_1$
18: end function
```

For the case of parallel querying, the acquisition function maximization is performed when a worker becomes available. To evaluate (14), we need to sample $\tilde{f}_*$, which is determined through f_* and $\boldsymbol{f}_{\mathcal{Q}}$ as shown in (11). This can be easily performed through RFM. By calculating $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, m)$ for $(\boldsymbol{x}, m) \in \mathcal{Q}$ with the sampled parameter $\boldsymbol{w}$, we can directly obtain a sample of $\boldsymbol{f}_{\mathcal{Q}}$. For f_*, we maximize $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, M)$ as in the sequential querying case. The algorithm of Parallel MF-MES is shown in Appendix D.3.

Throughout the paper, we use $I(f_*; f_{\boldsymbol{x}}^{(m)})$ as the information gain for brevity. $I(f_*; y_{\boldsymbol{x}}^{(m)})$, in which noisy observation $y_{\boldsymbol{x}}^{(m)}$ is contained, is also possible to use with the almost same procedure (for details, see Appendix C).

Although we mainly focus on the case that we only have the discrete fidelity level $m \in \{1, \dots, M\}$ as an "ordinal scale", several studies consider the setting in which a fidelity can be defined as a point z in a continuous "fidelity feature" (FF) space $\mathcal{Z}$ (Kandasamy et al., 2017). This setting is more restrictive because it requires additional side-information z which specifies a degree of fidelity, though this prior knowledge may be able to improve the accuracy. By introducing a kernel function in fidelity space $\mathcal{Z}$, our method can easily adapt to this setting (See appendix E).

5. Related Work

Multi-fidelity extension of BO has been widely studied. For example, (Huang et al., 2006; Lam et al., 2015; Picheny et al., 2013) extended the standard EI to the multi-fidelity setting. As with the usual EI, these are local measures of utility unlike the information-based approaches. *Gaussian process upper confidence bound* (GP-UCB) (Srinivas et al., 2010) is a popular approach in the single fidelity setting, and some studies proposed its multi-fidelity extensions. Kandasamy et al. (2016) proposed multi-fidelity GP-UCB for discrete fidelity $m = 1, \ldots, M$, and further, Kandasamy et al. (2017) proposed a similar UCB-based approach for the setting with the continuous fidelity space $\mathcal{Z}$. However, the UCB criterion has a trade-off parameter which balances exploit-exploration. In practice, this parameter needs to be carefully selected to achieve good performance. Another approach recently proposed in (Sen et al., 2018) is a multi-fidelity extension of a hierarchical space partitioning (Bubeck et al., 2011). However, this method assumes that the approximation error can be represented as a known function form of cost, and further, they associate fidelity with the depth of hierarchical tree, but the appropriateness of a specific choice of a pair of a point $\boldsymbol{x}$ and fidelity m is difficult to interpret.

Information-based BO has also been studied for the multi-fidelity setting, including *entropy search* (ES)-based (Swersky et al., 2013; Klein et al., 2017) and *predictive entropy search* (PES)-based (Zhang et al., 2017; McLeod et al., 2018) methods. Although these methods can measure global utility of the query without introducing any trade-off parameter, they inherit the computational difficulty of the original ES and PES, which consider the entropy of $p(\boldsymbol{x}_*)$, where $\boldsymbol{x}_* := \operatorname{argmax}_{\boldsymbol{x}} f(\boldsymbol{x})$ is the optimal solution. PES mitigates computational difficulty by using 1) the symmetric property of the mutual information, and 2) several assumptions which simplify involved densities. However, integral with respect to $\boldsymbol{x}_*$ is still necessary though the dimension of $\boldsymbol{x}_*$ can be high, and the complicated approximation procedure including *expectation propagation* (Minka, 2001) is required. Further, an additional assumption about inter-fidelity differences are required in the case of (Zhang et al., 2017). Song et al. (2018) proposed another information-based approach, which separates phases of the low-fidelity exploration and the highest fidelity optimization. However, the transition of these phases are controlled by a hyper-parameter which is necessary to set appropriately beforehand.

Another approach incorporating a measure of global utility is *knowledge gradient* (KG)-based methods (Poloczek et al., 2017; Wu & Frazier, 2017). This approach evaluates the max gain of predictive mean $\max_{\boldsymbol{x} \in \mathcal{X}} \mu_{\boldsymbol{x}}^{(M)}$. In particular, misoKG (Poloczek et al., 2017) deals with the discrete fidelity case. However, the acquisition function evaluation requires the expected value of the maximum of the mean function $\mathbb{E}[\max_{\boldsymbol{x}' \in \mathcal{X}} \mu_{\boldsymbol{x}'}^{(M)}]$ after adding $y_{\boldsymbol{x}}^{(m)}$ into training set, meaning that the maximization of the acquisition function is defined as a nested optimization. Although a variety of computational techniques have been studied for KG, this nested optimization process is highly cumbersome to implement and computationally expensive.

In contrast, our MF-MES is based on much simpler computations compared with existing information-based methods and other measures of global utility. Original MES calculates the entropy by representing a conditional distribution of $f_{\boldsymbol{x}}$ given f_* as a truncated normal distribution. As we saw in Section 3.1, for the information gain from a lower fidelity, the truncated normal approach is not applicable anymore because lower fidelity functions $f_{\boldsymbol{x}}^{(m)}$ for $m = 1, \ldots, M-1$ are not truncated for a given f_*. We already show that equations derived in Lemma 3.1 enables us to evaluate the entropy accurately with the only one dimensional additional numerical integration. For further acceleration of MES, Ru et al. (2018) proposed approximating the density of f_* and f given f_* by normal distributions, but reliability of these approximations are not clearly understood, and thus we do not employ in this paper.

The parallel extension of BO has been widely studied (e.g., Snoek et al., 2012; Desautels et al., 2014). As we described in Section 3.2, MFBO is typically asynchronous, while many of existing studies focus on the synchronous setting including PES-based parallel BO (Shah & Ghahramani, 2015). Several papers focus on the asynchronous setting (Kandasamy et al., 2018; Alvi et al., 2019), but these methods are difficult to apply to the multi-fidelity setting because they do not provide any criterion to select fidelity. To our knowledge, a KG-based method (Wu & Frazier, 2017) and BOHB (Falkner et al., 2018; Klein et al., 2020) are only parallel methods proposed for MFBO. However, the KG-based method is only for the synchronous setting, and further, it is only shown for the FF-based setting which is more restrictive as we described in the end of Section 4. BOHB combines BO and Hyperband (Li et al., 2018). Although asynchronous queries can be issued by A-BOHB (Klein et al., 2020), the BOHB-based methods focus on a more specific setting for the hyperparameter optimization of machine-learning algorithms. We also note that a parallel extension of MES has not been shown even for the single-fidelity setting. For possible sequential/parallel settings of MF-MES, a summary is shown in Appendix F.

6. Experiments

We evaluate effectiveness of MF-MES compared with other existing methods. To evaluate performance, we employed simple regret (SR) and inference regret (IR). SR is defined by $\max_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{x}}^{(M)} - \max_{\boldsymbol{x} \in \mathcal{X}_t^{(M)}} f_{\boldsymbol{x}}^{(M)}$, where $\mathcal{X}_t^{(M)}$ is a set of $\boldsymbol{x}$ for which a highest fidelity observation $y_{\boldsymbol{x}}^{(M)}$ is included in the training dataset at iteration t (note that $\mathcal{X}_t^{(M)}$ cannot be empty because highest fidelity observations must be included for the initial dataset as shown in Appendix G.1.1). SR indicates the error by the best point queried so far. IR is defined by $\max_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{x}}^{(M)} - f_{\hat{\boldsymbol{x}}_t}^{(M)}$, where $\hat{\boldsymbol{x}}_t := \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{X}} \mu_{\boldsymbol{x}}^{(M)}$ which is seen as the recommendation from the model at iteration t. If IR is larger than SR at an iteration, we employed the value of SR as IR of that iteration for stable evaluation. For MF-GPR, we used SLFM in GP-based methods, unless otherwise noted. For the kernel function, we used Gaussian kernel with automatic relevance determination (ARD).

We used a synthetic function generated by MF-GPR, two benchmark functions, and a real-world dataset from materials science. The details of functions are described as follows.

GP-based Synthetic Function: We generated $d = 3$ dimensional synthetic functions through an SLFM model that has $M = 2$ fidelities. The sampling cost is set as $(\lambda^{(1)}, \lambda^{(2)}) = (1, 5)$.

Benchmark Functions: We used two benchmark functions called Styblinski-Tang ($d = 2, M = 2$), and HartMann6 ($d = 6, M = 3$). The sampling cost of Styblinski-Tang and HartMann6 are set as $(\lambda^{(1)}, \lambda^{(2)}) = (1, 5)$ and $(\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) = (1, 3, 5)$, respectively.

Material Data: As an example of practical applications, we applied our method to the parameter optimization of a simulation model in materials science. The task is to optimize $d = 2$ material parameters in the simulation model (Tsukada et al., 2014). The objective function is the discrepancy between the precipitate shape predicted by the model and one measured by an electron microscope. Based on the numerical accuracy of the simulation model, the number of fidelities and the relative sampling cost are set as $M = 3$ and $(\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) = (5, 10, 60)$, respectively. Unlike other functions, the candidate $\boldsymbol{x}$ is fixed beforehand in this dataset (so-called the pooled setting). Each fidelity has 62,500 candidate points.

The experiments on the GP-based synthetic function were performed 100 times (10 different initialization for each one of 10 generated functions). The other benchmark functions and the material dataset were performed 10 times with different initialization. For further detail of the settings, see Appendix G.1.

6.1. Evaluation for Sequential Querying

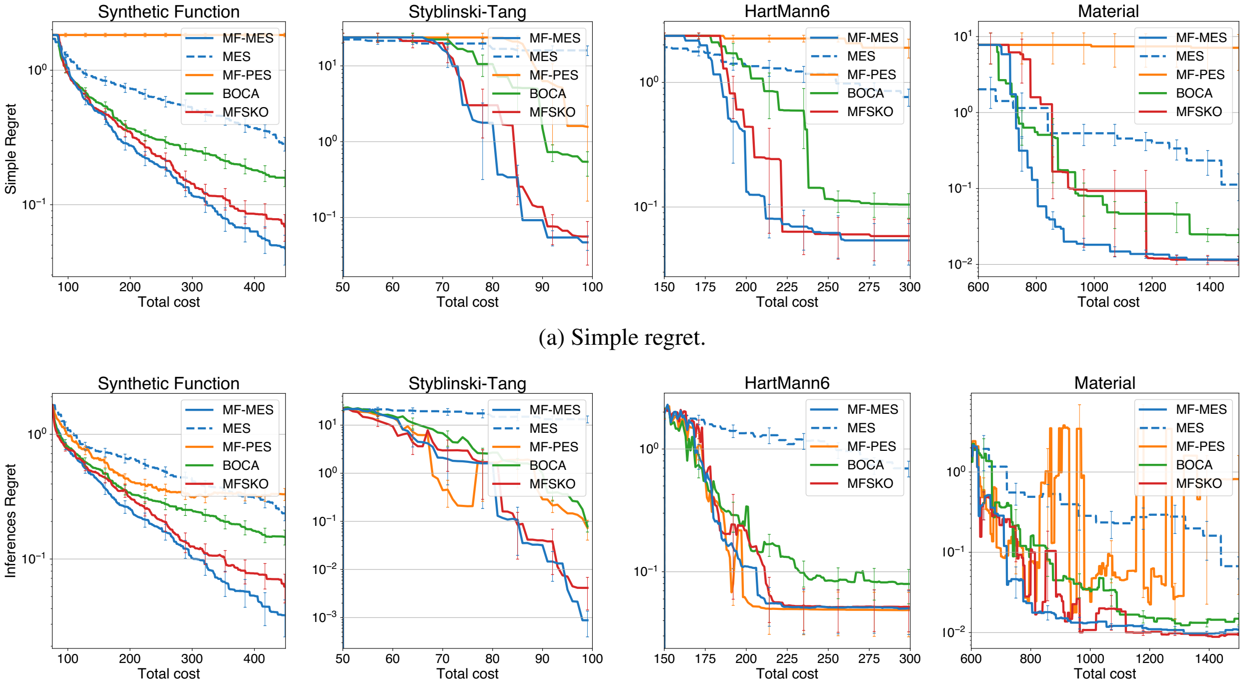
We first evaluate the performance for sequential querying. For comparison, we used MF-SKO (Huang et al., 2006), Bayesian optimization with continuous approximations (BOCA) (Kandasamy et al., 2017), and multi-fidelity PES (MF-PES) (Zhang et al., 2017). We also evaluated single fidelity MES which applied to the highest fidelity function $f^{(M)}(\boldsymbol{x})$. As we see in Section 5, misoKG is another measure of global utility for MFBO. However, we could not employ it as a baseline because it was not straightforward to modify the author implementation for fair comparison (e.g., changing the MF-GPR model), and creating efficient implementation from scratch is also extremely complicated (naïve implementation of KG can be prohibitively slow). Only BOCA employed the multi-task GPR (MT-GPR) model because the acquisition function assumes MT-GPR. For the sampling of f_* in MES and MF-MES, we employed the RFM-based approach described in Section 4, and sampled 10 f_*s at every iteration. In MF-PES, $\boldsymbol{x}_*$ was also sampled 10 times through RFM as suggested by (Hernández-Lobato et al., 2014).

Figure 2 shows SR and IR. In both of SR and IR, MF-MES decreased the regret faster than or comparable with all the other methods. The single-fidelity MES is relatively slow because it cannot use lower-fidelity functions, and we clearly see that MF-MES successfully accelerates MES. For SR of the GP-based synthetic, HartMann6 and material functions, MF-PES was slower than the others. We empirically observed that MF-PES sometime did not aggressively select the highest fidelity samples enough.

A possible reason is in an approximation employed by MF-PES which assumes $f_{\boldsymbol{x}}^{(m)} \leq f_{\boldsymbol{x}_*}^{(m)} + c$ for $m < M$, where c is a constant (see Zhang et al., 2017, for the detailed definition). However, even when $\boldsymbol{x}_*$ is given, this strict inequality relation does not hold obviously (note that $\boldsymbol{x}_*$ is the maximizer only when $m = M$), and we conjecture that the information gain from lower fidelity functions can be overly estimated because of this artificial truncation. In the material data, IR was slightly unstable which was caused by noisy observations contained in this real-world dataset. In particular, MF-PES largely fluctuated, and this would also be due to the lack of the highest fidelity samples as we mentioned above. We also evaluate computational time of the acquisition functions in Appendix G.2.

6.2. Evaluation for Parallel Querying

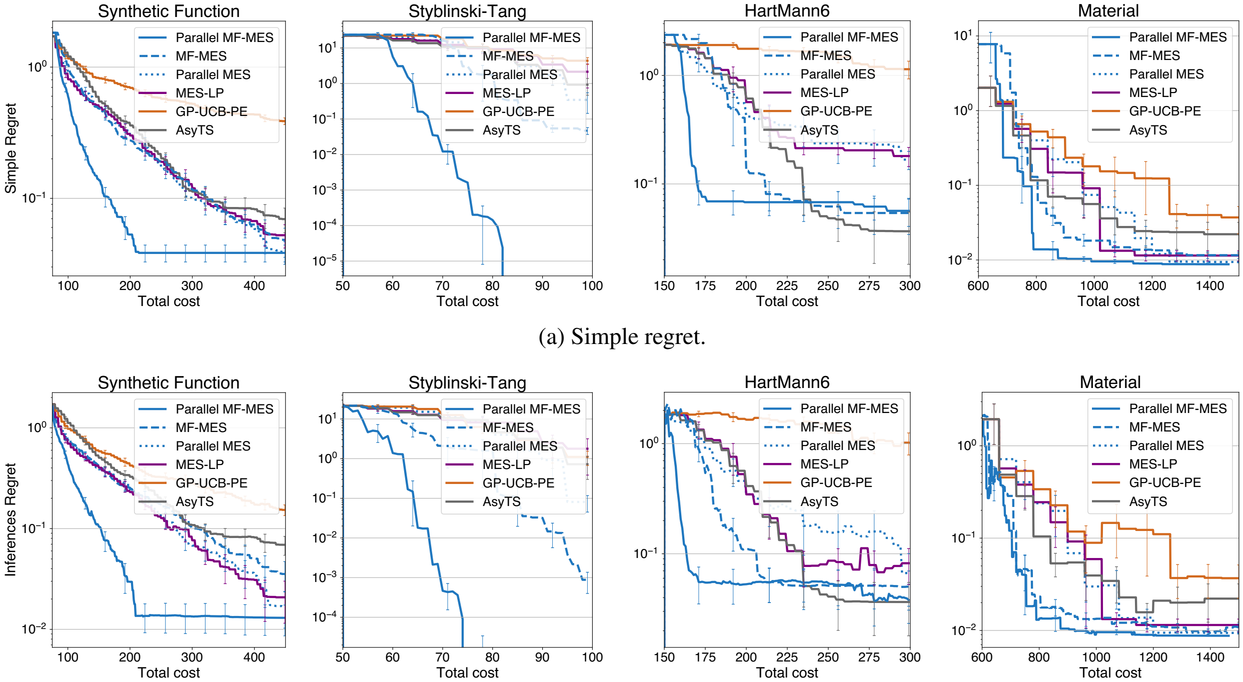
Next, we evaluate performance on parallel querying. For comparison, we used MES combined with local penalization (Gonzalez et al., 2016), denoted as MES-LP, Gaussian process upper confidence bound with pure exploration (GP-UCB-PE) (Contal et al., 2013), asynchronous parallel Thompson sampling (AsyTS) (Kandasamy et al., 2018). Here, we would like to note that no existing methods have been proposed for discrete fidelity parallel MFBO, to our knowledge, and extending existing methods to this setting



(a) Simple regret.

(b) Inference regret.

Figure 2: Performance comparison on sequential querying.



(a) Simple regret.

(b) Inference regret.

Figure 3: Performance comparison on parallel querying.

is not straightforward because of discreteness of fidelity levels. We also compare the performance of "sequential" MF-MES (which is same as "MF-MES" in Figure 2), and a parallel extension of single-fidelity MES (shown in Appendix D.4) as baselines. For the sampling of $\tilde{f}_*$ in Parallel MF-MES and Parallel MES, the number of samples are set 10 through RFM. The number of workers is set $q = 4$.

Figure 3 shows SR and IR. We see that parallel MF-MES substantially faster than sequential MF-MES and parallel MES. This indicates that parallel MF-MES succeeded in assigning workers across multiple fidelities. Compared with other methods, parallel MF-MES shows rapid or comparable convergence.

7. Conclusion

We propose a novel information-based multi-fidelity Bayesian optimization (MFBO). The acquisition function is defined through the information gain for the optimal value f_* of the highest fidelity function. We show that our method called MF-MES (multi-fidelity max-value entropy search) can be reduced to simple computations, which allows reliable evaluation of the entropy. For the asynchronous setting, which naturally arises in MFBO, we further propose parallelization of MF-MES and show that it is also easy to compute. We demonstrate effectiveness of MF-MES by using benchmark functions and a real-world materials science data. We showed the performance by total cost of function evaluations this time. However, wall-clock time is also often important in the multi-fidelity setting and the performance evaluation based on wall-clock time is one of our important future work.

ACKNOWLEDGMENTS

This work was supported by JST PRESTO (JPMJPR15NB, JPMJPR15N2 and JPMJPR16N6), Advanced Low Carbon Technology Research and Development Program (ALCA), MEXT KAKENHI (16H06538, 16H02866, 17H04694, 20H00601), MI2I project of JST Support Program for Starting Up Innovation Hub, JST CREST (JPMJCR1502), RIKEN Center for Advanced Intelligence Project, and RIKEN Junior Research Associate Program.

References

Alvi, A., Ru, B., Calliess, J.-P., Roberts, S., and Osborne, M. A. Asynchronous batch Bayesian optimisation with improved local penalisation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 253–262. PMLR, 09–15 Jun 2019.

Bhattacharjee, T., Mendis, C., Oh-ishi, K., Ohkubo, T., and Hono, K. The effect of ag and ca additions on the age hardening response of mg-zn alloys. *Materials Science and Engineering: A*, 575:231 – 240, 2013.

Bonilla, E. V., Chai, K. M., and Williams, C. Multi-task gaussian process prediction. In *Advances in Neural Information Processing Systems 20*, pp. 153–160. Curran Associates, Inc., 2008.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

Contal, E., Buffoni, D., Robicquet, A., and Vayatis, N. Parallel gaussian process optimization with upper confidence bound and pure exploration. In *Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases*, ECMLPKDD'13, pp. 225–240. Springer-Verlag, 2013.

Desautels, T., Krause, A., and Burdick, J. W. Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:4053–4103, 2014.

Falkner, S., Klein, A., and Hutter, F. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1437–1446. PMLR, 10–15 Jul 2018.

G, M. B. and Wilhelm, S. Moments calculation for the doubly truncated multivariate normal density, 2012.

Genton, M. G., Keyes, D. E., and Turkiyyah, G. Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, pp. 268–277, 2017.

Genz, A. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150, 1992.

Gonzalez, J., Dai, Z., Hennig, P., and Lawrence, N. Batch bayesian optimization via local penalization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pp. 648–657. PMLR, 2016.

Gumbel, E. J. *Statistics of Extremes*. Columbia University Press, 1958.

Hennig, P. and Schuler, C. J. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13:1809–1837, 2012.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems 27*, pp. 918–926. Curran Associates, Inc., 2014.

Huang, D., Allen, T., Notz, W., and Miler, R. Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32 (5):369–382, 2006.

Jones, D. R., Perttunen, C. D., and Stuckman, B. E. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1): 157–181, 1993.

Kandasamy, K., Dasarathy, G., Oliva, J., Schneider, J., and Póczos, B. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Advances in Neural Information Processing Systems 29*, pp. 1000–1008. Curran Associates, Inc., 2016.

Kandasamy, K., Dasarathy, G., Schneider, J., and Póczos, B. Multi-fidelity Bayesian optimisation with continuous approximations. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1799–1808, 2017.

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Poczos, B. Parallelised bayesian optimisation via Thompson sampling. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84, pp. 133–142. PMLR, 2018.

Kennedy, M. C. and O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

Klein, A., Falkner, S., Bartels, S., Hennig, P., and Hutter, F. Fast Bayesian optimization of machine learning hyperparameters on large datasets. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 528–536. PMLR, 2017.

Klein, A., Tiao, L. C., Lienart, T., Archambeau, C., and Seeger, M. Model-based asynchronous hyperparameter and neural architecture search, 2020.

Lam, R., Allaire, D. L., and Willcox, K. E. Multifidelity optimization using statistical surrogate modeling for non-hierarchical information sources. In *Proceedings of the 56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, pp. 0143. American Institute of Aeronautics and Astronautics, 2015.

Li, L., Jamieson, G. K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

McLeod, M., Osborne, M. A., and Roberts, S. J. Practical Bayesian optimization for variable cost objectives. *arXiv:1703.04335*, 2018.

Michalowicz, J. *Handbook of Differential Entropy*. Chapman and Hall/CRC, New York, 2014.

Minka, T. P. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann Publishers Inc., 2001.

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. Quantile-based optimization of noisy computer experiments with tunable precision. *Technometrics*, 55(1):2–13, 2013.

Poloczek, M., Wang, J., and Frazier, P. I. Multi-information source optimization. In *Advances in Neural Information Processing Systems 30*, pp. 4288–4298. Curran Associates, Inc., 2017.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. Curran Associates, Inc., 2008.

Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(54), 2017.

Ru, B., Osborne, M. A., Mcleod, M., and Granziol, D. Fast information-theoretic Bayesian optimisation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4384–4392. PMLR, 2018.

Sen, R., Kandasamy, K., and Shakkottai, S. Multi-fidelity black-box optimization with hierarchical partitions. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4538–4547. PMLR, 2018.

Shah, A. and Ghahramani, Z. Parallel predictive entropy search for batch global optimization of expensive objective functions. In *Advances in Neural Information Processing Systems 28*, pp. 3330–3338. Curran Associates, Inc., 2015.

Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25*, pp. 2951–2959. Curran Associates, Inc., 2012.

Song, J., Chen, Y., and Yue, Y. A general framework for multi-fidelity Bayesian optimization with gaussian processes. *arXiv:1811.00755*, 2018.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. 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*, pp. 1015–1022. Omnipress, 2010.

Swersky, K., Snoek, J., and Adams, R. P. Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems 26*, pp. 2004–2012. Curran Associates, Inc., 2013.

Teh, Y. W., Seeger, M. W., and Jordan, M. I. Semiparametric latent factor models. In *Proceedings of the 8th International Conference on Artificial Intelligence and Statistics*, 2005.

Tsukada, Y., Beniya, Y., and Koyama, T. Equilibrium shape of isolated precipitates in the α-mg phase. *Journal of Alloys and Compounds*, 603:65 – 74, 2014.

Wang, Z. and Jegelka, S. Max-value entropy search for efficient Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3627–3635. PMLR, 2017.

Wigley, P. B., Everitt, P. J., van den Hengel, A., Bastian, J. W., Sooriyabandara, M. A., McDonald, G. D., Hardman, K. S., Quinlivan, C. D., Manju, P., Kuhn, C. C. N., Petersen, I. R., Luiten, A. N., Hope, J. J., Robins, N. P., and Hush, M. R. Fast machine-learning online optimization of ultra-cold-atom experiments. *Scientific Reports*, 6:25890, 2016.

Wu, J. and Frazier, P. Continuous-fidelity Bayesian optimization with knowledge gradient. In *NIPS Workshop on Bayesian Optimization*, 2017.

Zhang, Y., Hoang, T. N., Low, B. K. H., and Kankanhalli, M. Information-based multifidelity Bayesian optimization. In *NIPS Workshop on Bayesian Optimization*, 2017.