Multi-objective Bayesian Optimization using Pareto-frontier Entropy

Shinya Suzuki [1] **Shion Takeno** [1 2] **Tomoyuki Tamura** [3 4] **Kazuki Shitara** [5 6] **Masayuki Karasuyama** [1 4 7]

Abstract

This paper studies an entropy-based multi-objective Bayesian optimization (MBO). Existing entropy-based MBO methods need complicated approximations to evaluate entropy or employ over-simplification that ignores trade-off among objectives. We propose a novel entropy-based MBO called *Pareto-frontier entropy search* (PFES), which is based on the information gain of *Pareto-frontier*. We show that our entropy evaluation can be reduced to a closed form whose computation is quite simple while capturing the trade-off relation in Pareto-frontier. We further propose an extension for the "decoupled" setting, in which each objective function can be observed separately, and show that the PFES-based approach derives a natural extension of the original acquisition function which can also be evaluated simply. Our numerical experiments show effectiveness of PFES through several benchmark datasets, and real-word datasets from materials science.

1. Introduction

This paper studies the black-box optimization problem with multiple objective functions. A variety of engineering problems require optimally designing multiple utility evaluations. For example, in materials design of the lithium-ion batteries, simultaneously maximizing ion-conductivity and stability is required for practical use. This type of problems can be formulated as jointly maximizing L unknown functions $f^1(\boldsymbol{x}), \ldots, f^L(\boldsymbol{x})$ on some input domain $\mathcal{X}$, which is called a *multi-objective optimization* (MOO) problem. MOO is often quite challenging because, typically, there does not exist any single optimal option due to the trade-off relation among different objectives. Further, as in the case of the single-objective black-box optimization, obtaining observations of each objective function is often highly expensive. For example, in the scientific experimental design such as synthesizing proteins, querying one observation can take more than a day. Since in MOO, the unique optimal point cannot be determined usually, a common approach is to search a set of *Pareto-optimal* points. For Pareto-optimal $\boldsymbol{f}_{\boldsymbol{x}} \coloneqq (f^1(\boldsymbol{x}), \ldots, f^L(\boldsymbol{x}))^\top$, there should not exist an alternative $\boldsymbol{f}_{\boldsymbol{x}'}$ that improves all the objectives simultaneously, and *Pareto-frontier* $\mathcal{F}^*$ is defined as a set of Pareto-optimal $\boldsymbol{f}_{\boldsymbol{x}}$s (see Section 2 for the formal definition).

1.1. Related Work

The combination of scalarization and evolutionary computations have been quite popular (e.g., Knowles, 2006; Zhang et al., 2010) for MOO problems. In particular, ParEGO (Knowles, 2006) has been widely known for its outstanding performance. The scalarization approach transforms MOO into a single-objective problem by which the Pareto-optimal solutions can be obtained under the certain regularity conditions. However, acquisition functions for the transformed single-objective are expected to be sub-optimal. Although recently, some studies (Paria et al., 2018; Marban & Frazier, 2017) have explored extensions of scalarization for identifying a specific subset of Pareto-frontier, we focus on identifying the entire Pareto-frontier in this paper.

Extending acquisition functions of usual *Bayesian optimization* (BO) has been a popular direction in MOO studies. An extension of standard *expected improvement* (EI) considers increase of Pareto hyper-volume (Emmerich, 2005), which we call *expected hyper-volume improvement* (EHI). Further, Shah & Ghahramani (2016) extended EHI to correlated objectives. Although EI is a widely accepted criterion, it measures the local utility only. *Upper confidence bound* (UCB) is another well-known acquisition function for BO (Srinivas et al., 2010). SMSego (Pon-

[1]Department of Computer Science, Nagoya Institute of Technology, Aichi, Japan [2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan [3]Department of Physical Science and Engineering, Nagoya Institute of Technology, Aichi, Japan [4]Center for Materials Research by Information Integration, National Institute for Material Science, Ibaraki, Japan [5]Joining and Welding Research Institute, Osaka University, Osaka, Japan [6]Nanostructures Research Laboratory, Japan Fine Ceramics Center, Aichi, Japan [7]PRESTO, Japan Science and Technology Agency, Saitama, Japan. Correspondence to: Masayuki Karasuyama <karasuyama@nitech.ac.jp>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

weiser et al., 2008) is one of UCB based approaches to MOO that optimistically evaluates the hyper-volume. PAL and ϵ-PAL (Zuluaga et al., 2013; 2016) are another UCB approaches in which a confidence interval based evaluation of Pareto-frontier is proposed. Shilton et al. (2018) evaluate the distance between a querying point and Pareto-frontier for defining a UCB criterion. A common difficulty of UCB approach is its hyper-parameter that balances the effect of the uncertainty term. Although there often exist theoretical suggestions for determining the hyper-parameter, careful tuning is necessary in practice since the suggested values usually contain unknown constants.

Campigotto et al. (2014) considers *uncertainty sampling* for directly modeling Pareto-frontier as a function. Although the simplest uncertainty sampling only measures local uncertainty at a querying point, global uncertainty measures have also been studied. SUR (Picheny, 2015) considers the expected decrease of *probability of improvement* (PI) as a measure of uncertainty reduction. However, SUR is computationally extremely expensive, because PI after a querying point is added to the training set is integrated over the entire $\mathcal{X}$, which severely limits scalability for the input space dimension.

In this paper, we particularly focus on the *information-theoretic* approach, which has been successful in single-objective BO (Hennig & Schuler, 2012; Hernández-Lobato et al., 2014; Wang & Jegelka, 2017). A seminal work of this direction for MOO is *predictive entropy search for multi-objective optimization* (PESMO), which defines an acquisition function through the entropy of a set of Pareto-optimal $\boldsymbol{x}$ (Hernandez-Lobato et al., 2016). They showed that the entropy-based acquisition function can achieve the superior performance compared with other types of criteria. However, PESMO employs an approximation based on *expectation propagation* (EP) (Minka, 2001) because the direct evaluation of their entropy is computationally intractable. In their EP, a non-Gaussian density is replaced with a Gaussian density, and it is difficult to show accuracy and reliability of this replacement. Further, the resulting calculation of the acquisition function is extremely complicated. On the other hand, Belakaria et al. (2019) proposed to use the entropy of the max-values of each dimension $l = 1, \dots, L$, called *max-value entropy search for multi-objective optimization* (MESMO). This drastically simplifies the calculation, but obviously, $\boldsymbol{f}_{\boldsymbol{x}} \in \mathcal{F}^*$ that does not have any values near the maximum of each axis is not preferred by this criterion. This means that trade-off relations among objectives, which are often essential for MOO problems, cannot be captured. For the relation with existing information-based methods, we further discuss in Section 5.

1.2. Contributions

We propose another entropy-based Bayesian MOO, called *Pareto-frontier entropy search* (PFES). We consider the entropy of the Pareto-frontier $\mathcal{F}^*$, defined in the space of the objective functions $\boldsymbol{f}_{\boldsymbol{x}}$, unlike PESMO that considers the entropy of the Pareto-optimal $\boldsymbol{x}$. By inheriting the advantage of the entropy-based approach, PFES provides a measure of global utility without any trade-off parameter. Under a few common conditions in entropy-based methods, we show that our acquisition function can be expressed as a closed form by using a cell-based partitioning of the output space. Although a naïve cell partitioning can generate a large number of cells for $L \geq 3$, we show an efficient computation by using a partitioning technique used in the *Pareto hyper-volume* computation. As a result, compared with PESMO, PFES provides a reliable evaluation of the entropy with much simpler computations. On the other hand, since the entire Pareto-frontier $\mathcal{F}^*$ is considered in the entropy calculation, PFES can capture trade-off relations in Pareto-frontier, which are ignored by MESMO.

We also discuss the *decoupled setting*, which was first introduced by (Hernandez-Lobato et al., 2016) as an extension of PESMO. This scenario assumes that each one of objective functions can be observed individually. Since observing all the objective functions simultaneously can cause huge cost, the decoupled setting evaluates only one of objectives at every iteration. Although this setting has not been widely studied, this can be highly important in practice. In particular, for search problems in scientific fields such as materials science and bio-engineering, multiple properties of objects (e.g., crystals, compounds, and proteins) can often be investigated separately by performing different experimental measurements or simulation-computations. In the battery material example, conductivity and stability can be evaluated through two independent physical simulations. Other directions of examples are also suggested by (Hernandez-Lobato et al., 2016), such as in robotics and design of a low calorie cookie (Solnik et al., 2017). However, the existing PESMO-based acquisition function, derived by naïvely decomposing the original acquisition function, was not fully justified in a sense of the entropy. We show that our PFES can be simply extended to the decoupled setting by considering the entropy of the marginal density without introducing any additional approximation.

Our numerical experiments show effectiveness of PFES through synthetic functions and real-world datasets from materials science.

2. Preliminary

We consider the *multi-objective optimization* (MOO) problem which maximizes $L \geq 2$ objective functions $f^l : \mathcal{X} \to \mathbb{R}$ for $l = 1, \ldots, L$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is an input domain. Let $\boldsymbol{f}_{\boldsymbol{x}} \coloneqq (f^1_{\boldsymbol{x}}, \ldots, f^L_{\boldsymbol{x}})^\top$, where $f^l_{\boldsymbol{x}} \coloneqq f^l(\boldsymbol{x})$. The optimal solution of MOO is usually defined by *Pareto optimality*. For a pair of $\boldsymbol{f}_{\boldsymbol{x}}$ and $\boldsymbol{f}_{\boldsymbol{x}'}$, if $f^l_{\boldsymbol{x}} \geq f^l_{\boldsymbol{x}'}$ for all $l \in \{1, \ldots, L\}$ with at least one of the inequalities being strict, we say "$\boldsymbol{f}_{\boldsymbol{x}}$ dominates $\boldsymbol{f}_{\boldsymbol{x}'}$" and the relation is denoted as $\boldsymbol{f}_{\boldsymbol{x}} \succ \boldsymbol{f}_{\boldsymbol{x}'}$. If $\boldsymbol{f}_{\boldsymbol{x}}$ is not dominated by any other $\boldsymbol{f}_{\boldsymbol{x}'}$ in the domain, $\boldsymbol{f}_{\boldsymbol{x}}$ is called Pareto-optimal. *Pareto-frontier* $\mathcal{F}^*$ is a set of Pareto-optimal $\boldsymbol{f}_{\boldsymbol{x}}$ which is written as $\mathcal{F}^* \coloneqq \{\boldsymbol{f}_{\boldsymbol{x}} \in \mathcal{F}_{\mathcal{X}} \mid \boldsymbol{f}_{\boldsymbol{x}'} \not\succ \boldsymbol{f}_{\boldsymbol{x}}, \forall \boldsymbol{f}_{\boldsymbol{x}'} \in \mathcal{F}_{\mathcal{X}}\}$, where $\mathcal{F}_{\mathcal{X}} \coloneqq \{\boldsymbol{f}_{\boldsymbol{x}} \in \mathbb{R}^L \mid \forall \boldsymbol{x} \in \mathcal{X}\}$. Although the Pareto-optimal points can be infinite, most strategies aim at finding a small subset of them that approximate the true $\mathcal{F}^*$ with sufficient accuracy.

Following the standard formulation of *Bayesian optimization* (BO), we model the objective function by Gaussian process regression (GPR). An observation for the l-th objective value of $\boldsymbol{x}_i$ is assumed to be $y^l_i = f^l_{\boldsymbol{x}_i} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$. The training dataset is written as $\mathcal{D} \coloneqq \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, where $\boldsymbol{y}_i = (y^1_i, \ldots, y^L_i)^\top$. Independent L GPRs are applied to each dimension with a kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$. By setting prior mean as 0, the predictive mean and variance of the l-th GPR are $\mu_l(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^\top \left(\boldsymbol{K} + \sigma^2_{\text{noise}}\boldsymbol{I}\right)^{-1} \boldsymbol{y}^l$, and $\sigma^2_l(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^\top \left(\boldsymbol{K} + \sigma^2_{\text{noise}}\boldsymbol{I}\right)^{-1} \boldsymbol{k}(\boldsymbol{x})$, where $\boldsymbol{k}(\boldsymbol{x}) \coloneqq (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n))^\top$, $\boldsymbol{y}^l \coloneqq (y^l_1, \ldots, y^l_n)^\top$, and $\boldsymbol{K}$ is the kernel matrix in which the i, j-element is defined by $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. We also define $\boldsymbol{\mu}(\boldsymbol{x}) \coloneqq (\mu_1(\boldsymbol{x}), \ldots, \mu_L(\boldsymbol{x}))^\top$ and $\boldsymbol{\sigma}(\boldsymbol{x}) \coloneqq (\sigma_1(\boldsymbol{x}), \ldots, \sigma_L(\boldsymbol{x}))^\top$. Although we assume L GPRs are independent throughout the paper, the extension for incorporating correlation among objectives are discussed in Appendix C.

3. Pareto-frontier Entropy Search for Multi-Objective Optimization

We propose a novel information-theoretic approach to multi-objective BO (MBO). Our method, called *Pareto-frontier entropy search* (PFES), considers maximizing the information gain for Pareto-frontier $\mathcal{F}^*$. With a slight abuse of notation, we write $\boldsymbol{f} \preceq \mathcal{F}^*$ when $\boldsymbol{f} \in \mathbb{R}^L$ is dominated by or equal to at least one of $\mathcal{F}^*$. The intuition behind Pareto-frontier entropy is shown in Fig. 1. The information gain can be evaluated by the mutual information between $\boldsymbol{f}_{\boldsymbol{x}}$ and Pareto-frontier $\mathcal{F}^*$, which we approximate as follows:

$$
\begin{aligned}
&I(\mathcal{F}^*; \boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}) \\
\approx &H[p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D})] - \mathbb{E}_{\mathcal{F}^*}\left[H[p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)]\right],
\end{aligned} \tag{1}
$$

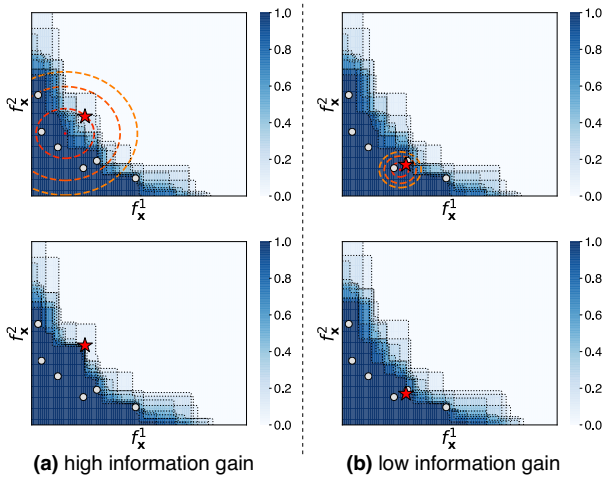


Figure 1: Illustrative examples of Pareto-frontier entropy. The blue heatmap represents $p(\boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ estimated by 10 Pareto-frontiers sampled from Gaussian process, and the white points are the observed data. The predictive distribution for $\boldsymbol{f}_{\boldsymbol{x}}$ is illustrated by the nested red circles in the above two plots. Suppose that the red star point is a sample generated by each predictive distribution, and the bottom two plots show $p(\boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ after adding the red star point into the training dataset. (a) Since the predictive distribution is on around the Pareto-frontier, the mutual information between $\boldsymbol{f}_{\boldsymbol{x}}$ and $\mathcal{F}^*$ is high. Therefore, when a sample is obtained from the predictive distribution, uncertainty of the Pareto-frontier is drastically reduced in the bottom plot. (b) Since the predictive distribution has low variance and is not particularly close to the Pareto-frontier, the mutual information is low in this case. Even if a sample is obtained from the predictive distribution, uncertainty of the Pareto-frontier is not largely changed.

where $H[\cdot]$ is the differential entropy. In the second term, we regard the conditional distribution $\boldsymbol{f}_{\boldsymbol{x}}$ given $\mathcal{F}^*$ as $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$, i.e., conditioning $\boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*$ only on the given $\boldsymbol{x}$ rather than requiring it for $\forall \boldsymbol{x} \in \mathcal{X}$. Note that the same simplification has been employed by most of existing state-of-the-art information-theoretic BO algorithms including well-known *predictive entropy search* (PES) and *max-value entropy search* (MES) proposed by Hernández-Lobato et al. (2014) and Wang & Jegelka (2017), respectively. Since the superior performance of these methods compared with other approaches has been shown, we also employ this simplification.

3.1. Acquisition Function

Since the first term in (1) is the simple L-dimensional Gaussian entropy, it can be analytically calculated. For the expectation over $\mathcal{F}^*$ in the second term, we use the Monte Carlo estimation. By sampling Pareto-frontier $\mathcal{F}^*$ from the

current GPR model, our acquisition function is written as follows:

$$a(\boldsymbol{x}) = H[p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D})] - \frac{1}{|\mathrm{PF}|} \sum_{\mathcal{F}^* \in \mathrm{PF}} H[p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)], \quad (2)$$

where PF is a set of sampled Pareto-frontier $\mathcal{F}^*$. We discuss the detail of the sampling procedure in Section 3.2.1. Although the entropy in the second term of (2) is still complicated at a glance, we show a tractable closed form of it by using a hyper-rectangle based partitioning of the dominated region.

Since the condition $\boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*$ indicates that $\boldsymbol{f}_{\boldsymbol{x}}$ must be dominated by or equal to at least one of Pareto-frontier $\mathcal{F}^*$, the density $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ is defined as a truncated distribution of the unconditional $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D})$ which is the predictive distribution of GPR, i.e., the independent multi-variate normal distribution. We call this distribution *Pareto-frontier truncated normal distribution* (PFTN). Figure 2 (a) and (b) illustrate the densities before and after the truncation, respectively. The density of PFTN $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ is written as

$$p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*) = \begin{cases} \frac{1}{Z} p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}) & \text{if } \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*, \\ 0 & \text{otherwise}, \end{cases}$$

where $Z = \int_{\boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*} p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}) \mathrm{d}\boldsymbol{f}_{\boldsymbol{x}}$ is a normalization constant. Let $\mathcal{F} \coloneqq \{\boldsymbol{f} \in \mathbb{R}^L \mid \boldsymbol{f} \preceq \mathcal{F}^*\}$ be the dominated region, and $M \in \mathbb{N}$ be the number of hyper-rectangles, called *cells*, by which the region $\mathcal{F}$ can be disjointly constructed as illustrated by Figure 2 (c). In other words, we can write $\mathcal{F} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \ldots \cup \mathcal{C}_M$, where the m-th cell $\mathcal{C}_m$ is defined by $(\ell_m^1, u_m^1] \times (\ell_m^2, u_m^2] \times \ldots \times (\ell_m^L, u_m^L]$. Note that this partitioning is created from $\mathcal{F}^*$ which is generated by the current GPR model (not from the observed data $\{\boldsymbol{y}_i\}_{i=1}^n$).

Let $\underline{\alpha}_{m,l} \coloneqq (\ell_m^l - \mu_l(\boldsymbol{x}))/\sigma_l(\boldsymbol{x})$, $\bar{\alpha}_{m,l} \coloneqq (u_m^l - \mu_l(\boldsymbol{x}))/\sigma_l(\boldsymbol{x})$, $Z_{ml} \coloneqq \Phi(\bar{\alpha}_{m,l}) - \Phi(\underline{\alpha}_{m,l})$, and $Z_m \coloneqq \prod_{l=1}^L Z_{ml}$, where Φ is the standard Gaussian cumulative distribution function (CDF). For the entropy in the second term of (2), the cell-based decomposition of the dominated region derives the following theorem (the proof is in Appendix A):

Theorem 3.1. *For L independent GPRs, the entropy of PFTN $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ is given by*

$$H[p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)] = \log\left((\sqrt{2\pi e})^L Z \prod_{l=1}^L \sigma_l(\boldsymbol{x})\right) + \sum_{m=1}^M \frac{Z_m}{Z} \sum_{l=1}^L \Gamma_{ml}, \quad (3)$$

where $\Gamma_{ml} \coloneqq (\underline{\alpha}_{m,l}\phi(\underline{\alpha}_{m,l}) - \bar{\alpha}_{m,l}\phi(\bar{\alpha}_{m,l}))/(2Z_{ml})$ with standard Gaussian probability density function (PDF) ϕ, and

$$Z = \sum_{m=1}^M \prod_{l=1}^L \int_{\ell_m^l}^{u_m^l} p(f_{\boldsymbol{x}}^l \mid \mathcal{D}) \mathrm{d}f_{\boldsymbol{x}}^l = \sum_{m=1}^M Z_m. \quad (4)$$

This entropy is a simple function of the predictive distribution of GPR at $\boldsymbol{x}$ and the Gaussian PDF/CDF functions. Thus, we can easily evaluate (3) if the cell-based partitioning is available. For the procedure of the partitioning, we discuss in Section 3.2.2.

3.2. Computation of Pareto-frontier Entropy

Suppose that we already have the predictive distribution of $\boldsymbol{x}$, i.e., $\boldsymbol{\mu}(\boldsymbol{x})$ and $\boldsymbol{\sigma}(\boldsymbol{x})$, a set of sampled Pareto-frontier PF, and a set of cells $\{\mathcal{C}_m\}_{m=1}^M$. Then, the normalization constant Z (4) is calculated by $O(ML)$, and the acquisition function (2) can also be obtained by $O(ML)$. We here describe the sampling procedure of $\mathcal{F}^*$, and the cell partitioning of the dominated region.

3.2.1. Sampling Pareto-frontier

PFES first needs to sample a set of Pareto-frontier $\mathcal{F}^*$. For this step, we follow an approach proposed by the existing information-based MBO (Hernandez-Lobato et al., 2016). They employed *random feature map* (RFM) (Rahimi & Recht, 2008) to approximate the current GPR by a Bayesian linear model $\boldsymbol{w}_l^\top \phi(\boldsymbol{x})$, where $\boldsymbol{w}_l \in \mathbb{R}^D$ is a parameter vector for the l-th objective and $\phi : \mathcal{X} \to \mathbb{R}^D$ is a pre-defined basis vector. By generating $\boldsymbol{w}_l$ from the posterior, we can sample a "function" $\boldsymbol{w}_l^\top \phi(\boldsymbol{x})$ with computational cost $O(D^3)$, and D is typically less than 1000. A sample of $\mathcal{F}^*$ can be obtained through solving MOO on $\boldsymbol{w}_l^\top \phi(\boldsymbol{x})$ for $l = 1, \ldots, L$. Since the objective $\boldsymbol{w}_l^\top \phi(\boldsymbol{x})$ can be easily evaluated for any $\boldsymbol{x}$, general MOO algorithms such as NSGA-II (Deb et al., 2002) is applicable. It has been empirically shown that entropy-based approaches are robust with respect to the number of this sampling (e.g., Wang & Jegelka, 2017), and usually only the small number of samples are used (e.g., 10). In the later experiments, we evaluate sensitivity of performance to this setting.

3.2.2. Partitioning of Dominated Region

For the generated Pareto-frontier, we need to construct a set of cells $\{\mathcal{C}_m\}_{m=1}^M$. A similar cell-based decomposition has been performed by existing MBO methods such as the well-known expected improvement-based method (Shah & Ghahramani, 2016) in a slightly different context. Although Shah & Ghahramani (2016) employed a naïve grid-based partitioning, which produces $O(|\mathcal{F}^*|^L)$ cells, this may cause large computational cost, particularly when $L > 2$. We show a method for the partitioning by which the number of cells can be drastically reduced compared with

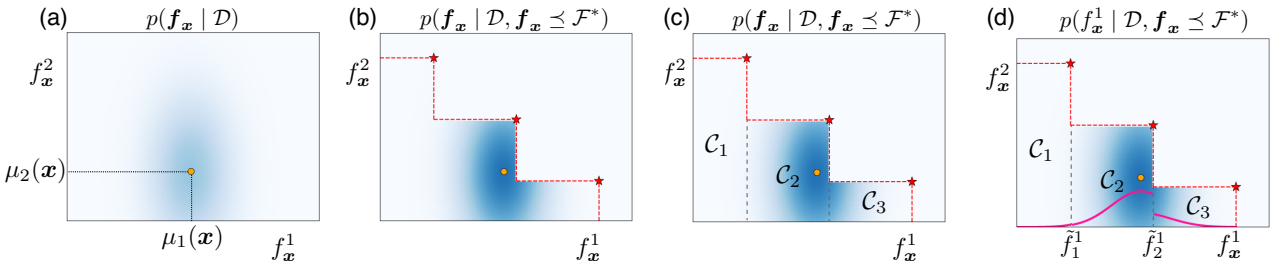


Figure 2: A schematic illustration of truncation by Pareto-frontier. (a) Original predictive distribution of two GPRs in the output space. (b) Predictive distribution truncated by Pareto-frontier, which results in PFTN. All $\boldsymbol{f}_{\boldsymbol{x}}$ should be dominated by the given Pareto-frontier (red stars). (c) Rectangle-based partitioning for the entropy evaluation. The entropy of PFTN is evaluated by decomposing the dominated region into rectangles called *cells* ($\mathcal{C}_1, \mathcal{C}_2$, and $\mathcal{C}_3$ in the plot). (d) Marginal density $p(f^1_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ considered in the decoupled setting (the solid pink line).

this naïve approach.

For $L = 2$, which is the most common setting in MOO, there exists a decomposition with $M = |\mathcal{F}^*|$ as we can clearly see in Figure 2 (c). Most of MOO algorithms (such as NSGA-II, used for generating $\mathcal{F}^*$) can explicitly specify the maximum number of $|\mathcal{F}^*|$ beforehand. This value is typically set as at most a few hundreds (Deb & Jain, 2014) even for a large L more than 10, which does not commonly occur in real-world MOO problems. We also empirically observed that a small $|\mathcal{F}^*|$ is sufficient to capture the trade-off relation among objectives. In our later experiments, we set $|\mathcal{F}^*| = 50$ by following an existing information-based approach (Hernandez-Lobato et al., 2016).

For $L > 2$, the simple partitioning like Figure 2 (c) is not applicable. To produce a smaller number of cells, we propose to use techniques in the *Pareto hyper-volume* computation. Pareto hyper-volume is defined by the volume of the region dominated by Pareto-frontier, which is widely used as an evaluation measure of MOO. Therefore, many studies have been devoted to its efficient computation mainly by decomposing the region into as few cells as possible (Couckuyt et al., 2014). Through this decomposition, we can obtain a partitioning as a by-product of the hyper-volume computation. For example, quick hyper-volume (QHV) (Russo & Francisco, 2014) is one of well-known methods which recursively calculates the volume by partitioning the region with a quick-sort like divide-and-conquer procedure. Under a few assumptions on the distribution of $\mathcal{F}^*$, QHV takes $O(L|\mathcal{F}^*|^{1+\epsilon}\log^{L-2}|\mathcal{F}^*|)$ time for the average case with high probability, where $\epsilon > 0$ is an arbitrary small constant (see Russo & Francisco, 2014, for the detail). Since the hyper-volume computation is still actively studied, more advanced algorithms are also applicable if it produces rectangle regions as a by-product of the algorithm. In this paper, we employ QHV because of its efficiency and simplicity.

4. Extension to Decoupled Setting

In the previous section, we assumed that all the objectives are observed simultaneously, to which we refer as the *coupled setting*. In contrast, the *decoupled setting* assumes that each one of objectives can be separately observed. Although this setting has not been widely studied in the MBO literature, this can be a significant problem setting particularly in the case that the sampling cost of each objective is highly expensive, because then, observing all the objectives every time may cause large amount of waste-of-cost. Further, we also focus on the fact that observation costs are often different among multiple objective functions, and introduce a *cost-sensitive* acquisition function into the decoupled setting.

In this setting, we need to determine a pair of an input $\boldsymbol{x}$ and an objective function index $l \in \{1, \ldots, L\}$ to be observed. PFES can provide a natural criterion for this purpose by considering the mutual information between $\mathcal{F}^*$ and the l-th objective $I(\mathcal{F}^*; f^l_{\boldsymbol{x}})$. We define the following cost-sensitive acquisition function:

$$a(\boldsymbol{x}, l) = \frac{1}{\lambda_l}\Big\{H[p(f^l_{\boldsymbol{x}} \mid \mathcal{D})] - \frac{1}{|\mathrm{PF}|}\sum_{\mathcal{F}^* \in \mathrm{PF}} H[p(f^l_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)]\Big\}, \tag{5}$$

where $\lambda_l > 0$ is the observation cost of the l-th objective which is assumed to be known beforehand. A pair of $\boldsymbol{x}$ and l to be queried can be determined by $\mathrm{argmax}_{\boldsymbol{x},l}\, a(\boldsymbol{x}, l)$. Here again, the first term of (5) is easy to calculate. We derive an efficient computation for the entropy in the second term. Figure 2 (d) shows an illustration of the density $p(f^l_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ in the second term.

Define $S \coloneqq |\mathcal{F}^*|$ as the number of the Pareto optimal points, and $\tilde{f}^l_1, \ldots, \tilde{f}^l_{S_l}$ for $S_l \leq S$ as a sequence ascendingly sorted by the l-th dimension of $\forall \boldsymbol{f}_{\boldsymbol{x}} \in \mathcal{F}^*$ in which duplicated values are eliminated. For $\forall m \in \{1, \ldots, M\}$,

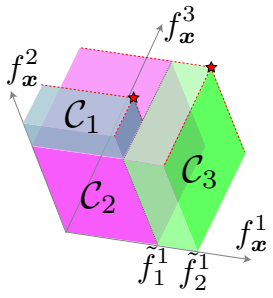


Figure 3: An example of partitioning for decoupled setting. In this case, $\mathcal{M}(l,s)$ for $l=1$ are $\mathcal{M}(1,0)=\{1,2\}$, and $\mathcal{M}(1,1)=\{3\}$.

we assume that there exists $s \in \{0,\dots,S_l\}$ such that $(\ell^l_m, u^l_m] = (\tilde{f}^l_s, \tilde{f}^l_{s+1}]$ (this assumption is just for notational simplicity, and we can create the cells in such a way that this condition is satisfied). The marginal density of PFTN $p(f^l_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ depends on the interval $(\tilde{f}^l_s, \tilde{f}^l_{s+1}]$ that $f^l_{\boldsymbol{x}}$ exists. Let $\mathcal{M}(l,s) := \{m \mid (\ell^l_m, u^l_m] = (\tilde{f}^l_s, \tilde{f}^l_{s+1}]\}$ be the index set of $\mathcal{C}_m$ in which the l-th dimension is equal to $(\tilde{f}^l_s, \tilde{f}^l_{s+1}]$ as illustrated in Figure 3, and $s^{(f)}_l \in \{0,\dots,S_l-1\}$ be the index s such that $f \in (\tilde{f}^l_s, \tilde{f}^l_{s+1}]$, where $\tilde{f}^l_0 := -\infty$.

Using independence of the objectives, we derive the following theorem (the proof is in Appendix B):

Theorem 4.1. *For L independent GPRs, the entropy of $p(f^l_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ is given by*

$$H[p(f^l_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)] = - \sum_{s=0}^{S_l-1} \frac{\sum_{m\in\mathcal{M}(l,s)} Z_m}{Z} \left(\log \frac{\sum_{m\in\mathcal{M}(l,s)} Z_m}{Z\sqrt{2\pi e}\sigma_l(\boldsymbol{x})\tilde{Z}_{sl}} - \tilde{\Gamma}_{sl} \right), \tag{6}$$

where $\tilde{Z}_{sl} := \Phi(\tilde{\alpha}_{s+1,l}) - \Phi(\tilde{\alpha}_{s,l})$ *with* $\tilde{\alpha}_{s,l} := (\tilde{f}^l_s - \mu_l(\boldsymbol{x}))/\sigma_l(\boldsymbol{x})$, *and* $\tilde{\Gamma}_{sl} := (\tilde{\alpha}_{s,l}\phi(\tilde{\alpha}_{s,l}) - \tilde{\alpha}_{s+1,l}\phi(\tilde{\alpha}_{s+1,l}))/(2\tilde{Z}_{sl})$.

As shown in this theorem, even for the decoupled case, we obtain a closed form representation of the entropy. Although the equation may look complicated, this can be easily calculated from the predictive distribution if the cell-based partitioning is given.

The computation for the acquisition function of the decoupled setting is similar to the coupled case. We first sample $\mathcal{F}^*$ from the current GPR model, and then, creating the cell partitioning for each sampled Pareto-frontier. The partitioning shown in Figure 3 can also be created from the QHV partitioning. For each interval $(\tilde{f}^l_s, \tilde{f}^l_{s+1}]$, if a cell $\mathcal{C}$ created by QHV contains this interval, we extract a sub-cell $\mathcal{C}' \subseteq \mathcal{C}$ in which only the interval of the l-th dimension of $\mathcal{C}$ is replaced with $(\tilde{f}^l_s, \tilde{f}^l_{s+1}]$. This procedure increases the total number of cells at most $|\mathcal{F}^*|$ times which we usually set a small value (50 in this paper). After the partitioning, for a given predictive distribution of GPR, the acquisition function (5) can be simply calculated with $O(M|\mathcal{F}^*|L)$ by using (6).

5. Relation with Other Information-theoretic Approaches

For MBO, two other information-theoretic approaches, called *Predictive entropy search for multi-objective optimization* (PESMO) (Hernandez-Lobato et al., 2016) and *Max-value entropy search for multi-objective optimization* (MESMO) (Belakaria et al., 2019), have been proposed. Herein, we discuss relation of our PFES with those existing methods.

PESMO considers the entropy of a set of Pareto optimal $\boldsymbol{x}$, defined as Pareto set $\mathcal{X}^*$. PESMO first samples $\mathcal{X}^*$ from the current model, and consider the entropy of $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \mathcal{X}^*)$. However, unlike the case of PFES, the entropy is not directly reduced to a closed form. An *expectation propagation* (EP) (Minka, 2001) based approximation results in that the each dimension of $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \mathcal{X}^*)$ is represented by an independent Gaussian distribution, whose accuracy and appropriateness are not clarified. Further, the computational procedure of this approximation is highly complicated. By contrast, in our PFES, PFTN $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ and its entropy is computationally tractable without approximations, and thus, the dependent relation in this density is incorporated into the acquisition function evaluation. From Figure 2 (b), we can clearly see that $p(\boldsymbol{f}_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ can have dependent relation among $\boldsymbol{f}_{\boldsymbol{x}}$ nevertheless the original GPR is assumed to be independent. Although PESMO is the only method that is applicable to the decoupled setting, the acquisition function is derived by simply decomposing the information gain of the coupled setting, and an interpretation of what the decomposed value represents has not been explicitly shown. On the other hand, in PFES, the entropy of the conditional density $p(f^l_{\boldsymbol{x}} \mid \mathcal{D}, \boldsymbol{f}_{\boldsymbol{x}} \preceq \mathcal{F}^*)$ is directly derived as shown in (6).

MESMO is another information-based MOO that uses the entropy of the max-values of each dimension $l=1,\dots,L$. MESMO is inspired by *max-value entropy search* (MES) of single-objective BO (Wang & Jegelka, 2017) that considers the entropy of the optimal output $\max_{\boldsymbol{x}} f(\boldsymbol{x})$. This approach drastically simplifies the calculation, but obviously in MOO, Pareto-frontier is not constructed only by the max-value of each axis, and thus, $\boldsymbol{f}_{\boldsymbol{x}} \in \mathcal{F}^*$ that does not have values near the max-values is not preferred by this criterion. For example, although the red star point in Figure 1 (a) is not the maximum in the both axes, this point

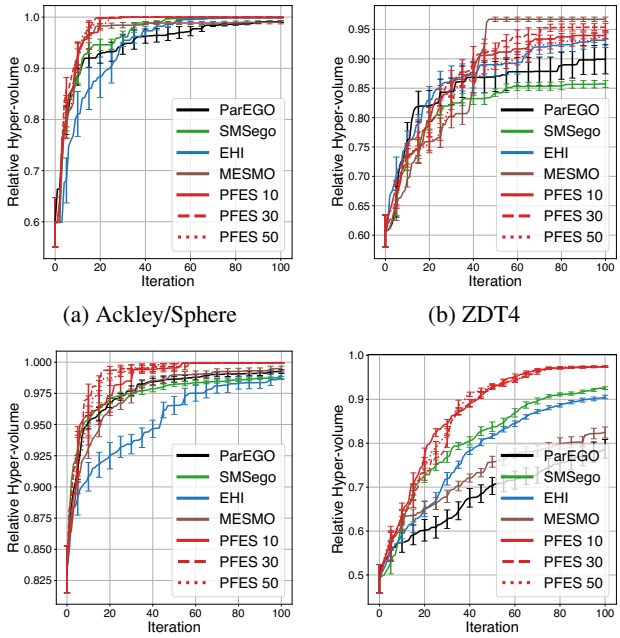
largely improves the Pareto-frontier created by the already observed points (white circles). In fact, in a sense of the entropy evaluation defined by (3), PFES cannot be worse than MESMO because MESMO replaced it with an approximation. Although MESMO claims convergence by using a well-known R_2 indicator (Hansen & Jaszkiewicz, 1998) like criterion, the MESMO convergence analysis is difficult to interpret. The analysis first calculates $f^l_{\boldsymbol{x}^*} - f^j_{\boldsymbol{x}_t}$, where $\boldsymbol{x}^*$ is a Pareto-optimal solution and $\boldsymbol{x}_t$ is the selected point at the t-th iteration. Note that this value can be negative, because $f^l_{\boldsymbol{x}^*}$ is not necessarily the maximum of f_l. To define a cumulative evaluation through iterations, this difference for each j is accumulated respectively, and after that, a norm is taken over the cumulative values of all objectives. Although each cumulative value before taking the norm can be negative, an interpretation about this cumulative value is not clarified.

6. Experiments

We compared PFES with ParEGO, EHI, SMSego, and MESMO. To evaluate performance, we used the hyper-volume of the region dominated by Pareto-frontier, which is a standard evaluation measure in MOO. For the kernel function in all the methods, we employed the Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2/(2\sigma^2))$. The samplings of $\mathcal{F}^*$ in PFES and $\mathcal{X}^*$ in MESMO, which we call *Pareto sampling*, were performed 10 times, respectively. For the cell partitioning of PFES, we used the QHV algorithm (Russo & Francisco, 2014). For the acquisition function maximization of all methods, we used the DIRECT algorithm (Jones et al., 1993). The performance is evaluated by the hyper-volume created by already observed instances relative to the optimal hyper-volume, which we call *relative hyper-volume* (RHV). The other experimental settings are shown in Appendix F. Because of implementation and computational complexity issues, we could not perform comparison with PESMO and SUR while they provide the global measures of utility for MOO (the author implementation was not compatible with our environment). We believe that the above compared methods are currently widely used, and thus, would be sufficient as the baseline to verify the performance of PFES.

6.1. Benchmark Functions

We first used benchmark functions which have continuous domain $\mathcal{X}$. Each experiment run 10 times with a different set of initial observations which were randomly selected 5 points. Here, we consider the coupled setting. For Pareto sampling, NSGA-II was applied to functions generated from RFM with 500 basis functions, and we set the maximum size of Pareto set as 50 by following (Hernandez-Lobato et al., 2016). The results on four MOO problems



(a) Ackley/Sphere (b) ZDT4

(c) DTLZ3 (d) DTLZ4

Figure 4: Performance comparison on benchmark problems (average and standard error of 10 runs).

are shown in Figure 4. In the figure, (a) Ackley/Sphere is created by combining two single objective benchmark functions $L = 2$ with $d = 2$ (Surjanovic & Bingham, 2013), and (b) - (d) are from well-known MOO benchmark functions (Huband et al., 2006). ZDT4 has two objectives $L = 2$ and the input dimension is $d = 4$. DTLZ3 and 4 have four objectives $L = 4$ and the input dimension is $d = 6$. To calculate RHV, the optimal hyper-volume is estimated by applying NSGA-II to the true objective function. Here, in PFES, we evaluate the three settings of the number of samplings $|\mathrm{PF}| = 10, 30$, and 50.

Figure 4 shows the results. We see that PFES (any of 10, 30 or 50) achieved the fastest convergence for Ackley/Sphere, DTLZ3, and DTLZ4, and for DTLZ3, PFES was roughly the second best among the compared methods. The three settings of $|\mathrm{PF}|$ showed similar behaviors, suggesting that the performance dependence of PFES on $|\mathrm{PF}|$ is small. MESMO showed the faster convergence on ZDT4, while for the two MOO benchmark functions DTLZ3 and DTLZ4, it was relatively slow. Among the three MOO benchmark functions, only ZDT4 has the "convex" dominated region $\mathcal{F}$, while DTLZ3 and DTLZ4 have the "concave" $\mathcal{F}$ (see Huband et al., 2006, for the detail). The stronger trade-off relation exists in the concave case, for which MESMO failed to improve RHV rapidly.

We also examined the computational time for the acquisition function evaluation on DTLZ4 which has the largest

Table 1: Computational time for acquisition function evaluation for 100 points on DTLZ4.

(a) Comparison of five methods (sec)

ParEGO	SMSego	EHI	MESMO	PFES
0.53	6.32	317.55	59.90	62.36

(b) Details of PFES (sec, except for # cells)

Total	RFM	NSGA-II	QHV	Entropy	#cells
62.36	0.11	59.85	1.28	1.13	647.86

output dimension $L = 4$ in our four benchmark functions. We randomly selected 50 training instances and also randomly selected 100 candidate $\boldsymbol{x}$ to evaluate the acquisition functions. The average of 10 runs of this procedure is shown in Table 1 (a). ParEGO and SMSego are fast because their acquisition functions are simple. Although EHI took relatively long time, this is mainly because we employed the naïve cell partitioning described in (Shah & Ghahramani, 2016). MESMO and PFES are similar computational times. Table 1 (b) shows the detailed elapsed time in PFES. We see that the most of time was spent by NSGA-II in this case. The amount of QHV and the entropy calculation is quite small, and #cells is less than 1,000 even in this $L = 4$ problem which is a middle-large sized MOO problem because a problem $L \geq 4$ is sometimes called a "many-objective" problem in the context of MOO (Chand & Wagner, 2015). Since MESMO also employed NSGAII (Belakaria et al., 2019), MESMO and PFES showed similar result. We further report with other settings in Appendix E.

6.2. Decoupled Setting with Materials Data

For evaluating the decoupled acquisition function, we used two real-world datasets from *computational materials science*. In this field, efficient exploration of materials is strongly demanded because accurate physical simulations are often computationally extremely expensive, in which simulations taking more than several days are common. The task is to explore crystal structures achieving high ion-conductivity and stability (i.e., $L = 2$), which are desirable properties for battery materials. For these datasets, $\mathcal{X}$ is a pre-defined discrete set, meaning that we have a fixed number of candidates (the pooled setting). Details of the two datasets, called Bi_2O_3 and LLTO, are as follows:

Bi_2O_3 The size of candidates is $|\mathcal{X}| = 335$, generated by the composition $Bi_{1-x-y-z}Er_xNb_yW_zO_{48+y+3/2z}$. The input is the three dimensional space defined by x, y, and z.

LLTO The size of candidates is $|\mathcal{X}| = 1119$, generated by the crystal called Perovskite type $La_{2/3-x}Li_{3x}TiO_3$ for $x = 0.11$. In each candidate, positions of each one of atoms are permuted. The 2185 dimensional feature vector $\boldsymbol{x}$ is created through relative three dimensional positions of the atoms. Note that although this dataset has the high dimensional input space, BO is feasible because $\mathcal{X}$ is the pre-defined discrete set.

The objective functions are ion-conductivity $f^1_{\boldsymbol{x}}$ and stability $f^2_{\boldsymbol{x}}$ (negative of the energy), which can be observed through physical simulation models, separately. The Bi2O3 and LLTO data are collected based on quantum- and classical- mechanics, respectively. In the both cases, ion-conductivity is more expensive because it requires time-consuming simulations for observing dynamics of the ion. Here, we examine the two cost settings $(\lambda_1, \lambda_2) = (5, 1)$ and $(\lambda_1, \lambda_2) = (10, 1)$, based on the prior knowledge of the domain experts. In these datasets, PFES directly generated function values of GPR without RFM, from which the Pareto set can be easily sampled unlike the continuous input case. Each experiment run 10 times with a different set of initial observations which were randomly selected 5 points.

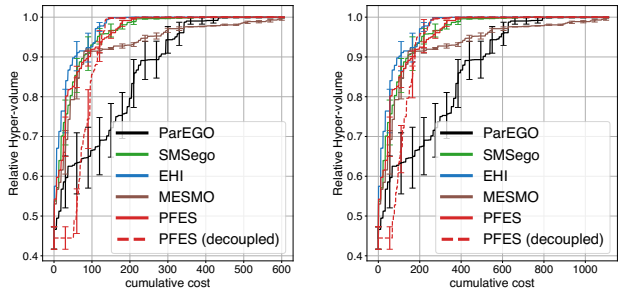
Figure 5 and 6 show the result. The horizontal axis of the figure is the sum of the observation cost. In Bi_2O_3, SMSego, EHI, PFES, and PFES (decoupled) showed relatively rapid convergence, while in LLTO, PFES (decoupled) reached the maximum first. Interestingly, for the both datasets, the increase of PFES (decoupled) was moderate compared with the other methods in the beginning, and it was accelerated at the middle of iterations. PFES (decoupled) starts sampling from low cost functions because in the beginning the amount of information from the two objectives are not largely different. After collecting cheaper information, PFES (decoupled) moves onto the expensive objectives and the faster improvement of RHV compared with the coupled PFES was finally observed in a sense of the total sampling cost.

7. Conclusion

We proposed *Pareto-frontier entropy search* (PFES) for multi-objective Bayesian optimization (MBO). We showed that the entropy of Pareto-frontier can be simply evaluated via sampling of Pareto-frontier and the cell-based partitioning. Further, we showed PFES for the decoupled setting through the marginalization, for which simple computations are also obtained. Our empirical evaluation on the benchmark functions and materials science data demonstrated effectiveness of our approach.

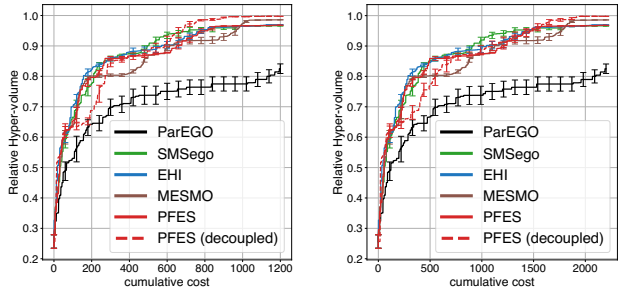
ACKNOWLEDGMENTS

This work was supported by MEXT KAKENHI 17H04694, 18K04700, JST PRESTO JPMJPR15N2, "Materials research by Information Integration" Initiative (MI^2I) project of the Support Program for Starting Up



(a) $(\lambda_1, \lambda_2) = (5, 1)$ (b) $(\lambda_1, \lambda_2) = (10, 1)$

Figure 5: RHV for Bi_2O_3.



(a) $(\lambda_1, \lambda_2) = (5, 1)$ (b) $(\lambda_1, \lambda_2) = (10, 1)$

Figure 6: RHV for LLTO.

Innovation Hub from JST, and RIKEN Junior Research Associate Program.

References

Belakaria, S., Deshwal, A., and Doppa, J. R. Max-value entropy search for multi-objective bayesian optimization. In *Advances in Neural Information Processing Systems 32*, pp. 7823–7833. Curran Associates, Inc., 2019.

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.

Campigotto, P., Passerini, A., and Battiti, R. Active learning of pareto fronts. *IEEE Transactions on Neural Networks and Learning Systems*, 25(3):506–519, 2014.

Chand, S. and Wagner, M. Evolutionary many-objective optimization: A quick-start guide. *Surveys in Operations Research and Management Science*, 20(2):35 – 42, 2015.

Couckuyt, I., Deschrijver, D., and Dhaene, T. Fast calculation of multiobjective probability of improvement and expected improvement criteria for pareto optimization. *Journal of Global Optimization*, 60(3):575–594, Nov 2014.

Deb, K. and Jain, H. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*, 18(4):577–601, 2014.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197, 2002.

Emmerich, M. T. M. Single- and multi-objective evolutionary design optimization assisted by Gaussian random field metamodels, 2005. PhD thesis, FB Informatik, University of Dortmund.

Genz, A. and Bretz, F. *Computation of Multivariate Normal and T Probabilities*. Springer Publishing Company, Incorporated, 2009.

Hansen, M. P. and Jaszkiewicz, A. Evaluating the quality of approximations to the non-dominated set, 1998. IMM Technical Report IMM-REP-1998-7.

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

Hernandez-Lobato, D., Hernandez-Lobato, J., Shah, A., and Adams, R. Predictive entropy search for multi-objective Bayesian optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 1492–1501. PMLR, 2016.

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.

Huband, S., Hingston, P., Barone, L., and While, L. A review of multiobjective test problems and a scalable test problem toolkit. *IEEE Transactions on Evolutionary Computation*, 10(5):477–506, 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.

Knowles, J. ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006.

Manjunath, B. and Wilhelm, S. Moments calculation for the double truncated multivariate normal density. *SSRN eLibrary*, 2009.

Marban, R. A. and Frazier, P. Multi-attribute Bayesian optimization under utility uncertainty. In *NIPS Workshop on Bayesian Optimization*, 2017.

Michalowicz, J. V., Nichols, J. M., and Bucholtz, F. *Handbook of differential entropy*. Chapman and Hall/CRC, 2013.

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

Paria, B., Kandasamy, K., and Pòczos, B. A flexible framework for multi-objective Bayesian optimization using random scalarizations. *arXiv:1805.12168*, 2018.

Picheny, V. Multiobjective optimization using gaussian process emulators via stepwise uncertainty reduction. *Statistics and Computing*, 25(6):1265–1280, 2015.

Ponweiser, W., Wagner, T., Biermann, D., and Vincze, M. Multiobjective optimization on a limited budget of evaluations using model-assisted S-metric selection. In *Parallel Problem Solving from Nature – PPSN X*, pp. 784–794. Springer Berlin Heidelberg, 2008.

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.

Russo, L. M. S. and Francisco, A. P. Quick hypervolume. *IEEE Transactions on Evolutionary Computation*, 18(4): 481–502, 2014.

Seeger, M., Teh, Y.-W., and Jordan, M. I. Semiparametric latent factor models. Technical report, Workshop on Artificial Intelligence and Statistics 10, 2004.

Shah, A. and Ghahramani, Z. Pareto frontier learning with expensive correlated objectives. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 1919–1927. JMLR.org, 2016.

Shilton, A., Rana, S., Gupta, S. K., and Venkatesh, S. Multi-target optimisation via Bayesian optimisation and linear programming. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 145–155, 2018.

Solnik, B., Golovin, D., Kochanski, G., Karro, J. E., Moitra, S., and Sculley, D. Bayesian optimization for a better dessert. In *Proceedings of the 2017 NIPS Workshop on Bayesian Optimization*, 2017.

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.

Surjanovic, S. and Bingham, D. Virtual library of simulation experiments: Test functions and datasets. Retrieved January, 2019, from `http://www.sfu.ca/~ssurjano`, 2013.

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.

Zhang, Q., Liu, W., Tsang, E., and Virginas, B. Expensive multiobjective optimization by MOEA/D with gaussian process model. *IEEE Transactions on Evolutionary Computation*, 14(3):456–474, 2010.

Zuluaga, M., Sergent, G., Krause, A., and Püschel, M. Active learning for multi-objective optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 462–470. PMLR, 2013.

Zuluaga, M., Krause, A., and Püschel, M. e-PAL: An active learning approach to the multi-objective optimization problem. *Journal of Machine Learning Research*, 17(104):1–32, 2016.