Filtering Bayesian optimization approach in weakly specified search space

Filtering Bayesian Optimization Approach in Weakly Specified Search Space

Vu Nguyen · Sunil Gupta · Santu Rana · Cheng Li · Svetha Venkatesh

Abstract Bayesian optimization (BO) has recently emerged as a powerful and flexible tool for hyper-parameter tuning and more generally for the efficient global optimization of expensive black-box functions. Systems implementing BO has successfully solved difficult problems in automatic design choices and machine learning hyper-parameters tunings. Many recent advances in the methodologies and theories underlying Bayesian optimization have extended the framework to new applications and provided greater insights into the behavior of these algorithms. Still, these established techniques always require a user-defined space to perform optimization. This pre-defined space specifies the ranges of hyper-parameter values. In many situations, however, it can be difficult to prescribe such spaces, as a prior knowledge is often unavailable. Setting these regions arbitrarily can lead to inefficient optimization - if a space is too large, we can miss the optimum with a limited budget, on the other hand, if a space is too small, it may not contain the optimum point that we want to get. The unknown search space problem is intractable to solve in practice. Therefore, in this paper, we narrow down to consider specifically the setting of "weakly specified" search space for Bayesian optimization. By weakly specified space, we mean that the pre-defined space is placed at a sufficiently good region so that the optimization can expand and reach to the optimum. However, this pre-defined space need not include the global optimum. We tackle this problem by proposing the filtering expansion strategy for Bayesian optimization. Our approach starts from the initial region and gradually expands the search space. We develop an efficient algorithm for this strategy and derive its regret bound. These theoretical results are complemented by an extensive set of experiments on benchmark functions and two real-world applications which demonstrate the benefits of our proposed approach.

Keywords Bayesian optimization · unknown search space · hyper-parameter tuning · experimental design

V. Nguyen, S. Gupta, S. Rana, C. Li, S. Venkatesh
Center for Pattern Recognition and Data Analytics (PRaDA)
Deakin University, Australia
E-mail: v.nguyen@deakin.edu.au

1 Introduction

Global optimization is fundamental to diverse real-world problems where parameter settings and design choices are pivotal - as an example, in algorithm hyper-parameter tuning [42,39] or engineering design [9,8,2]). This requires us to optimize a non-concave objective function using sequential and noisy observations. Critically, the objective functions are unknown and expensive to evaluate. The challenge is to find the maximum of such expensive objective functions in few sequential queries to minimize time and cost.

Bayesian optimization (BO) [16,13,12,33] is an approach to optimize these above expensive functions. BO finds a solution of an expensive black-box function $\mathbf{x}^* = \underset{\mathbf{x}\in\mathscr{X}}{\operatorname{argmax}} f(\mathbf{x})$ by making a series of evaluations $\mathbf{x}_1,...,\mathbf{x}_T$ of f such that the optimum of f is found in the fewest iterations. As more data are observed, BO finds the optimum by sequentially updating the model, typically through a Gaussian process (GP) [34]. Using this GP posterior, BO builds a surrogate model - known as acquisition function - to select a next point to evaluate.

Existing Bayesian optimization approaches are restricted to a pre-defined and fixed space of search wherein it is assumed to contain the global optimum. Unfortunately, setting these regions so that it encapsulates the global optimum is non-trivial and often done arbitrarily. The main reason is that in many situations specifying a search space $\mathscr{X}$ is hard for a new problem or where domain knowledge is limited. This remain a key challenge that hinders us from getting the best performance for global optimization.

To ensure that a global optimum is found, BO requires good coverage of $\mathscr{X}$ by sufficient evaluations, but as the space increases, the number of evaluations needed also increases accordingly. Given the same evaluation budget, it is harder to find the optimum in larger search spaces compared to smaller ones. For example, finding a missing aircraft would usually take longer if a search space is the whole ocean instead of only a small island. Therefore, a search space greatly influences the performance of Bayesian optimization and its judicious choice remains an open problem.

As solving a general unknown space in Bayesian optimization is intractable, we focus on a scenario in BO where the search space is "weakly known" by the domain experts [28]. Although this user-specified space may not contain the global optimum, it is specified in a sufficiently good region so that the optimization can expand and reach to the optimum location under limited evaluation budget. This assumption is essential for efficient optimization and to avoid the situations where the optimum location is unreachable from the initial region due to budgetary constraints.

To address the weakly specified problem for Bayesian optimization, we propose a filtering expansion strategy that starts from an initial region and gradually expands it to find the optimum. Our filtering expansion for Bayesian optimization (FBO) includes two steps. It first utilizes the posterior mean and variance of the GP to guide the selection of promising regions to explore. Next, within the expanded region, it maximizes the acquisition function to select the next point to evaluate. We derive the regret bound for convergence analysis of the model. We demonstrate the efficacy of our approach in expanding a search space by optimizing several benchmark functions and hyper-parameter tuning of multi-label classification algorithm. Next, we show the most compelling example which is the experimental design for the aeronautical alloy AA2050. Although BO [45,23,15] is the ideal method for optimizing advanced materials, it is not clear for the domain experts to define the right region for search [1]. Therefore, we make use of the weakly defined region by our metallur-

[1] a trivial range of is too large and defect the purpose of Bayesian optimization by easily exceeding the evaluation budget.

gist collaborators and show how our algorithm can be successfully employed for this alloy design. Our contributions are:

- Formulation of the novel method for a weakly specified space setting in Bayesian optimization;
- Derivation of the regret bound providing guarantee for convergence analysis under the expanding space setting;
- Demonstration of the proposed method for optimizing both benchmark functions and hyper-parameter tuning of machine learning algorithms;
- Application to advanced material optimization on the aeronautical alloy design where the region of search is only weakly specified.

We structure our paper as follows. In Sec. 2, we present the preliminary background on Bayesian optimization including Gaussian process and acquisition function. We also present the algorithm and illustrative example of Gaussian process and Bayesian optimization. Next, we discuss the impact of the search space toward optimization and the possible problem of unknown search space in Sec. 3. Then, we present our proposed framework in Sec. 4 with the algorithm and convergence analysis. In Sec. 5, we present the experiments on benchmark functions and real-world applications. Finally, we provide the discussion, conclusion and future work.

2 Bayesian Optimization

We use f to denote a black-box function for which we have no closed-form expression. Furthermore, this black-box function is expensive to evaluate. Perturbed evaluations of the form $y_i = f(\mathbf{x}_i) + \varepsilon_i$ are available, where we assume the perturbations to follow a Gaussian distribution, i.e. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Formally, let $f : \mathscr{X} \to \mathscr{R}$ be a well-behaved function defined on a subset $\mathscr{X} \subseteq \mathscr{R}^d$. Our goal is to solve the following global optimization problem

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathscr{X}}{\operatorname{argmax}} f(\mathbf{x}). \tag{1}$$

As in other kinds of optimization, Bayesian optimization aims to find the global optimum of the black-box function $f(x)$. What makes Bayesian optimization different from other procedures is that it constructs a probabilistic model for $f(x)$ and then exploits this model to make decisions about where in $\mathscr{X}$ to next evaluate the function, while integrating out uncertainty. The essential philosophy is to use all of the information available from previous evaluations of $f(x)$ and not simply rely on local gradient and Hessian approximations. This results in a procedure that can find the minimum of difficult non-convex functions with relatively few evaluations, at the cost of performing more computation to determine the next point to try [5,38,37,24]. We summarize the routine for BO in *Algorithm* .1.

There are three major choices that must be made when performing Bayesian optimization. First, one must select a prior over functions that will express assumptions about the function being optimized. For this we choose the Gaussian process prior, due to its flexibility and tractability. Second, we must choose an acquisition function, which is used to construct a utility function from the model posterior, allowing us to determine the next point to evaluate. Third, we need to specify the search space to perform optimization. For this requirement, we note that the existing BO approaches assume the argmax to be restricted to a bounded subset $\mathscr{X} \subset \mathscr{R}^d$ while our approach will flexibly open this bound.

2.1 Gaussian Process

Bayesian optimization reasons about f by building a Gaussian process through evaluations [34]. This flexible distribution allows us to associate a normally distributed random variable at every point in the continuous input space. Formally, the function is modelled by $f(\mathbf{x}) \sim \mathrm{GP}(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}'))$, where m is the mean and $k(\mathbf{x},\mathbf{x}')$ contains the covariance of any two observations. A popular choice for the covariance function is the squared exponential: $k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp\left[\frac{-(\mathbf{x}-\mathbf{x}')^2}{2l^2}\right]$ where l is the length scale and σ_f^2 is the output variance. The length scale defines the "region of influence" of a point within the parameter space that the influence of an observation decreases as one considers points farther away from this observation. The output variance σ_f^2 defines the expected deviation of the function output y away from its average value. In practice, we can standardize the output $y \sim \mathcal{N}(0,1)$ and set $\sigma_f^2 = 1$ for simplicity.

We get the predictive distribution for a new observation $\mathbf{x}'$ that also follows a Gaussian distribution [34] - its mean and variance are given by:

$$\mu\left(\mathbf{x}'\right) = \mathbf{k}(\mathbf{x}',X)\mathbf{K}(X,X)^{-1}\mathbf{y} \tag{2}$$

$$\sigma^2\left(\mathbf{x}'\right) = k(\mathbf{x}',\mathbf{x}') - \mathbf{k}(\mathbf{x}',X)\mathbf{K}(X,X)^{-1}\mathbf{k}(\mathbf{x}',X)^T \tag{3}$$

where $K(U,V)$ is a covariance matrix whose element (i,j) is calculated as $k_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)$ with $\mathbf{x}_i \in U$ and $\mathbf{x}_j \in V$.

Although we have analytic expressions, exact inference in GP is $\mathcal{O}(N^3)$ where N is the number of observations. This cost is due to the inversion of the covariance matrix. In practice, the Cholesky decomposition can be computed once and saved so that subsequent predictions are $\mathcal{O}(N^2)$. However, this Cholesky decomposition must be recomputed every time the kernel hyper-parameters changed, which usually happens at every iteration. For large data sets, or large function evaluation budgets in the Bayesian optimization setting, the cubic cost of exact inference is prohibitive and there have been many attempts at reducing this computational burden via approximation techniques, such as using sparse Gaussian process [32,21]. Alternative solution to Gaussian process, people have used Bayesian deep learning [40], deep neural network [39] or random forest [4]. We provide an illustrative example of Gaussian process in Fig. 1.

2.2 Acquisition functions

As the original function $f(x)$ is expensive to evaluate, we seek to replace it by the acquisition function $\alpha(x)$ which is cheaper. Built on a Gaussian process, this acquisition function determines a next point to evaluate. Therefore, instead of maximizing the original function, we maximize the acquisition function to select the next point

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathscr{X}}{\operatorname{argmax}}\,\alpha_t(\mathbf{x}).$$

In this auxiliary maximization problem, the objective is known and can be easily optimized by standard numerical techniques such as multi-start or DIRECT [19] which typically requires a fixed region $\mathscr{X}$. Although the acquisition functions can be constructed using various regression models, such as Gaussian process [34,21], random forest [17], neural networks

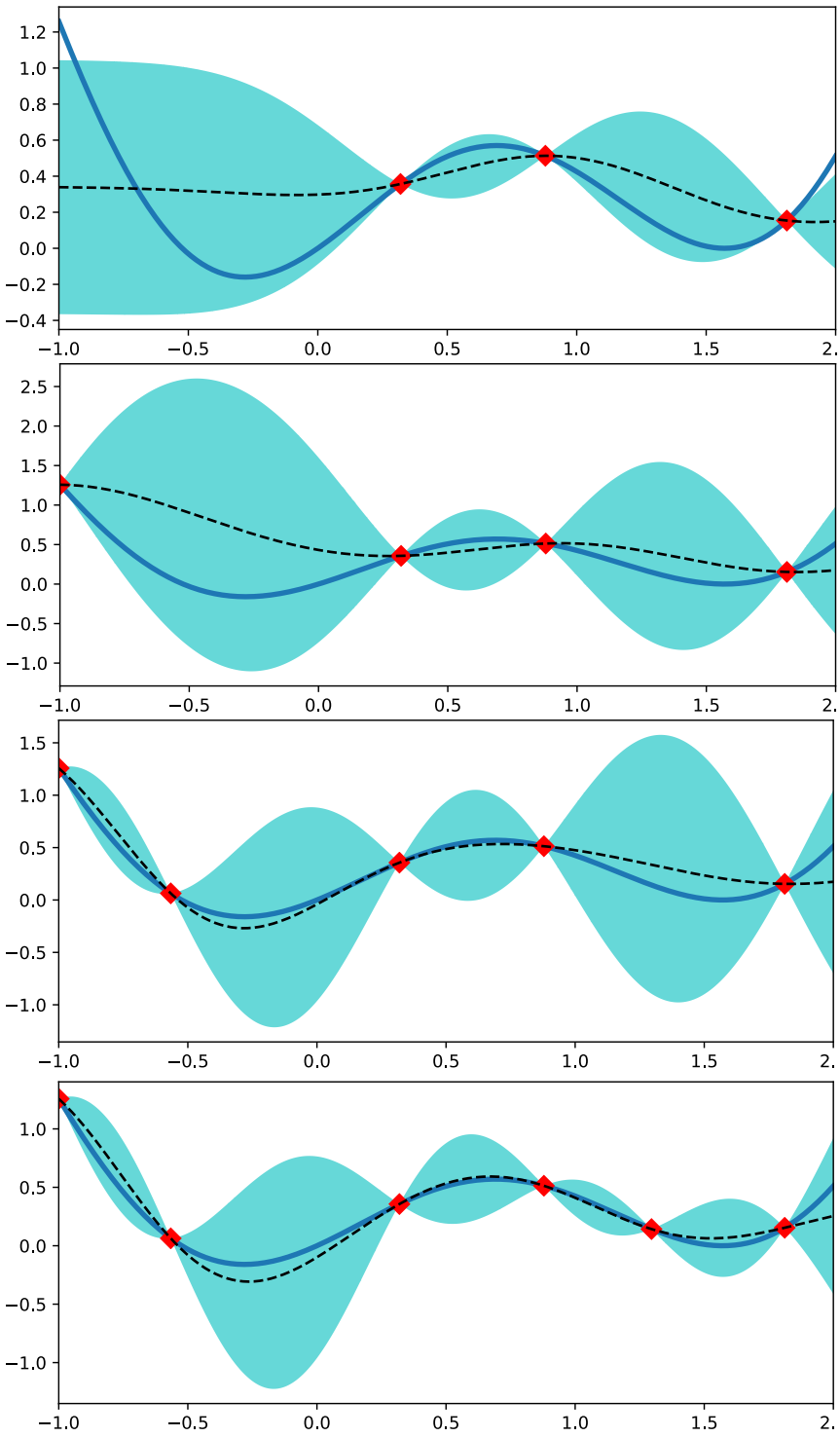


Fig. 1: Example of Gaussian Process and Bayesian Optimization. Red dots are the observations x. Blue line is the true function $y = f(x)$. Black dash line is the GP predictive mean $\mu(x)$. Blue shaded area is the GP predictive variance $\sigma(x)$. The GP predictive mean is more accurate and the GP predictive variance is reducing when we have more observations. (best viewed in color).

Algorithm .1 Bayesian Optimization.

Input: initial data $\mathscr{D}_0$, #iter T
1: **for** $t = 1$ to T **do**
2: Fit a GP from $\mathscr{D}_t$ to get the predictive mean $\mu(x)$ and predictive variance $\sigma(x)$.
3: Obtain $\mathbf{x}_t = \underset{x \in \mathscr{X}}{\operatorname{argmax}} \alpha_t(x, \mu, \sigma)$.
4: Evaluate the function $y_t = f(\mathbf{x}_t)$.
5: Augment the data $\mathscr{D}_t = \mathscr{D}_{t-1} \cup (\mathbf{x}_t, y_t)$.
6: **end for**
Output: $\mathbf{x}_{\max}, y_{\max}$

[39,40], in this paper we consider specifically the case of constructing the acquisition function $\alpha(\mathbf{x})$ from the posterior distribution of GP.

The acquisition functions are carefully designed to trade off between exploration of the search space and exploitation of current promising regions. Although many acquisition functions have been proposed [25,13,18,41,43,27], no single acquisition strategy provides the best performance over all problem. We review common acquisition functions: GP-UCB and EI in the following.

2.2.1 Gaussian process upper confidence bound (GP-UCB)

The GP-UCB [41] algorithm is defined as

$$\alpha^{\text{UCB}}(\mathbf{x}) = \mu(\mathbf{x}) + \sqrt{\beta}\sigma(\mathbf{x})$$

where β is a parameter to balance exploration and exploitation. There are theoretically motivated guidelines [41] for setting β to achieve sublinear regret.

2.2.2 Expected improvement (EI)

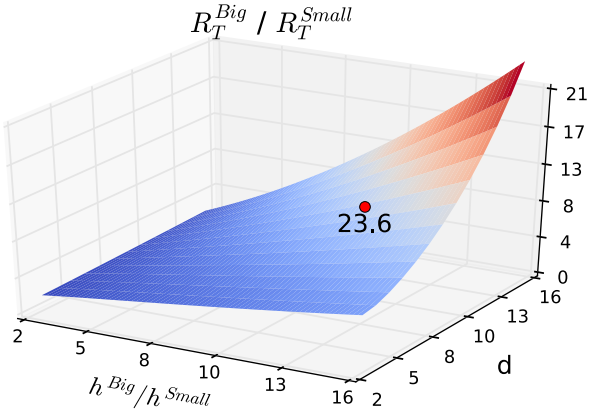
The expected improvement (EI) [25] measures the amount of improvement over the incumbent (e.g., the maximum value observed so far $y^+ = \max_{\forall y_i \in D_t} y_i$). First, we define the improvement function, denoted by $I^{\text{EI}}(x) = \max\{0, f(\mathbf{x}) - y^+\}$. Then, we take the expectation over the improvement function $\mathbb{E}\left[I^{\text{EI}}(\mathbf{x}, \theta)\right]$ which can be computed analytically. In particular, given $z = \frac{\mu(\mathbf{x}) - y^+}{\sigma(\mathbf{x})}$, the EI is computed as

$$\alpha^{\text{EI}}(\mathbf{x}) = \left[\mu(\mathbf{x}) - y^+\right]\Phi(z) + \sigma(\mathbf{x})\phi(z)$$

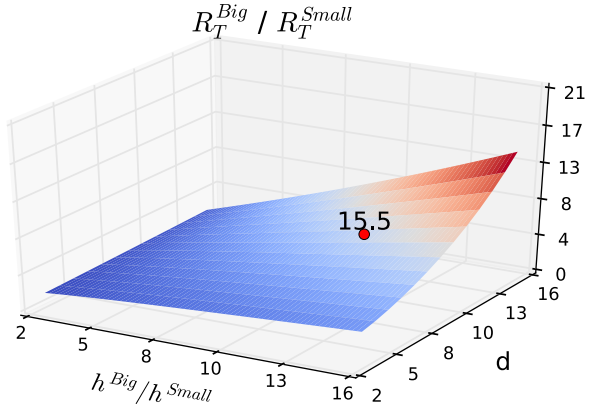
where Φ and ϕ are the normal c.d.f. and p.d.f. Recently the convergence rates have been proven for EI [6,44,30].

3 Impact of the Search Space to Bayesian Optimization

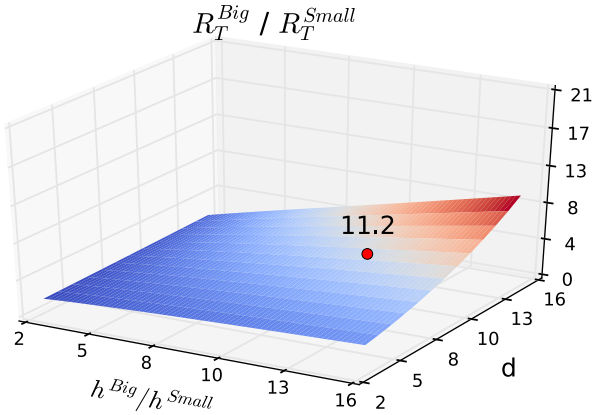
It is intuitive that performing optimization on a large space is more difficult and requires more effort than a small space given the same evaluation budget. To highlight this effect, we derive that a small space will have better (smaller) regret proportionally to its volume compared to a large space.



(a) $T = 5d$



(b) $T = 10d$



(c) $T = 20d$

Fig. 2: The regret bound is proportional to the space $V(X) = h^d$ (where h is a radius and d is a dimension) and the evaluation budget T. Let us consider a dimension $d = 10$ and radius ratio be $\frac{h^{\text{Big}}}{h^{\text{Small}}} = 12$. The regret bound of R_T^{Small} is better than R_T^{Big} with a factor of $23.6, 15.5$ and 11.2 as we increase the number of iterations $T = 5d, 10d$ and $20d$, respectively. Thus, given a small evaluation budget in BO, it is not efficient for optimization to define an arbitrarily large search space.

3.1 A search space influences the regret

Srinivas et al [41] has provided the sublinear regret bound for BO. However, the previous study has not highlighted the role of a search space. Therefore, in the following, we analyze the impact of search spaces (large vs small) toward the cumulative regret bound. We first follow [41] to define the maximum information gain γ_T. We then use Theorem 8 in [41] to derive a bound on γ_T, given a radius h s.t. the volume of the space $V(\mathscr{X}) \approx h^d$ (d being the dimension of the space)

$$\gamma_T \leq \max_{r=1,\ldots,T} T_* \log\left(r \times n_T/\sigma^2\right) + C(r,T,T_*)$$

where T is the number of iteration, $T_* \propto [\log(T \times n_T)]^d$, $n_T \propto 2V(\mathscr{X})\, T^d (\log T)$ and $C(r,T,T_*)$ is a function which will be zero by setting $r = T$. Then, we express the above equation by plugging T_*, n_T and set $r = T$ (as in [41]) to have

$$\gamma_T \leq \left[\log\left(2V(\mathscr{X})\, T^{d+1} (\log T)\right)\right]^{d+1}.$$

We consider two settings using a big and a small space defined by the volume with the radius h, s.t. $h^{\text{Small}} = \frac{h^{\text{Big}}}{\alpha}$ where $\alpha \geq 1$. We obtain the relation between the maximum information gain

$$\frac{\gamma_T^{\text{Big}}}{\gamma_T^{\text{Small}}} \propto \left[\frac{\log\left(T^{d+1}(\alpha h^{\text{Small}})^d \log T\right)}{\log\left(T^{d+1}(h^{\text{Small}})^d \log T\right)}\right]^{d+1} \geq 1. \tag{4}$$

Finally, we achieve the relation in the regret bound of the small vs the big space by utilizing the form of $R_T \overset{\Delta}{=} \sqrt{T\beta_T\gamma_T}$ [41]

$$\frac{R_T^{\text{Big}}}{R_T^{\text{Small}}} = \sqrt{\frac{\gamma_T^{\text{Big}}}{\gamma_T^{\text{Small}}}}. \tag{5}$$

In Eq. (5), we can interpret the regret bound relation that is proportional to the space (by the radius h and the dimension d). First, a larger space will have larger (worse) regret bound. For example, if $\frac{h^{\text{Big}}}{h^{\text{Small}}}$ increases, the fraction $\frac{R_T^{\text{Big}}}{R_T^{\text{Small}}}$ also increases accordingly - given that the other terms are fixed. Second, a smaller number of evaluation budget T will make the difference in the regrets more significant, i.e. $\frac{R_T^{\text{Big}}}{R_T^{\text{Small}}}$ is larger. Third, a higher dimension d makes the difference in the regrets larger almost exponentially in Eq. (4).

We plot this regret bound relations in Fig. 2 by varying the numbers of evaluations as $T = 5d, 10d$ and $20d$, respectively. Recall that the smaller is better for the regret, we show that the cumulative regret is reduced, i.e. $R_T^{\text{Big}} \geq R_T^{\text{Small}}$, when we let the space be smaller $h^{\text{Big}} \geq h^{\text{Small}}$. Given the fixed space and dimension, the regret is even worse if we have a smaller budget of evaluations T - the case in Bayesian optimization.

3.2 The problems of unknown search space

Since a search space is unknown and previous approaches do not have any provision to specify it properly. Instead, previous works specify them quite arbitrarily. Arbitrary setting these regions can lead to critical issues of either over-specifying or under-specifying. If we over-specify a search space by setting it too big (to make sure it contains the global optimum), the optimization is not efficient and can be intractable in high dimension: given the limited number of evaluations, as discussed in Sec. 3.1. On the other hand, if we under-specify a search space by setting it small, then the global optimum may lie outside the box and we again miss the optimum. Therefore, there is a trade-off between over- and under-specifying a search space in Bayesian optimization. To prevent these issues in BO, we narrow down the problem and consider incorporating the search space information where the pre-defined region is vaguely experienced by a domain expert and need not contain the global optimum.

3.3 Unbounded Bayesian optimization

Although determining the space for BO is a critical issue in practice, there is little research on this track due to the difficulty [35]. To the best of our knowledge, [36] is the first work to tackle the unbounded BO where a search space is unknown. The algorithm starts from an initial box and growing.

In particular, Shahriari et al [36] propose two strategies including volume doubling and regularization. The first approach of volume doubling is a heuristic method to expand a search space frequently as the optimization progresses. This approach requires a parameter specifying how often the expansion is occurred. The second approach is using regularization of a search space from a center. Basically, it is motivated by a regularized version for improvement policies, e.g., EI [25]. The authors [36] propose two choices of regularizing as quadratic (Q) and isotropic hinge-quadratic (H),

$$\xi_Q(\mathbf{x}) = (\mathbf{x}-\bar{\mathbf{x}})^T \operatorname{diag}\left(w^2\right)^{-1} (\mathbf{x}-\bar{\mathbf{x}})$$
$$\xi_H(\mathbf{x}) = \mathbb{I}\left[||\mathbf{x}-\bar{\mathbf{x}}||_2 > R\right] \left(\frac{||\mathbf{x}-\bar{\mathbf{x}}||_2 - R}{\beta R}\right)^2.$$

However, these regularizing approaches still suffer from three limitations. First, it requires additional parameters (R, β and w) which are sensitive and difficult to specify in practice. Second, they expand the search space equally to all directions from the mean location ($\bar{\mathbf{x}}$). This equal expansion makes a search space larger than required. Thus, it is slow to move toward the optimal region which can be far from the mean location. Finally, there is no theoretical guarantee on the efficiency of these optimization approaches [36].

4 Filtering Expansion Strategy for Bayesian Optimization in Weakly Specified Space

We consider the Bayesian optimization setting that a search space for optimization is hard to specify and only weakly specified due to practical situations, e.g. lacking expert knowledge or new domain application. By weakly specified space, we mean that this region is placed at a sufficiently good region so that the optimization can expand and reach to the optimum location under limited evaluation budget, but this initial region need not contain the global optimum. We describe most of the notations in Fig. 3 and Table 1 for clarity and ease of understanding.

Notation	Type	Meaning
σ_l^2	constant	length-scale for a kernel, e.g. squared exponential kernel
σ^2	constant	noise output variance (or measurement noise)
$\sigma^2(x)$,$\sigma(x)$	function	GP predictive variance function
$\mu(x)$	function	GP predictive mean function
d	constant	dimension of input feature x
$\mathscr{X} \in \mathscr{R}^d$	domain	a whole search space (unknown) $\forall x \in \mathscr{X}$
$\mathscr{X}_0 \in \mathscr{R}^d$	domain	an initial search space (weakly specified)
$\mathscr{X}_t \in \mathscr{R}^d$	domain	a search space at iteration t
$y = f(x)$	scalar	black-box function evaluation at x
f^*	scalar	global optimum, $f^* = \max_{\forall x \in \mathscr{X}} f(x)$
f_t^*	scalar	intermediate optimum within $\mathscr{X}_t$, $f_t^* = \max_{\forall x \in \mathscr{X}_t} f(x)$
r_t	scalar	regret at iteration t, $r_t = f^* - f(x_t)$
R_T	scalar	cumulative regret, $\sum r_t$
X_t	matrix	collection of input $X_t = [x_1, ...x_t]$
Y_t	vector	collection of outcome $Y_t = [y_1 ....y_t]$
T	scalar	number of iteration (evaluation budget)
$\mathscr{D}_t$	set	observation set upto iteration t,$\mathscr{D}_t = \{X_t, Y_t\}$
$\|\|x - x'\|\|^2$	function	$\sum_{i=1}^d (x_i - x_i')^2$

Table 1: Notation list.

4.1 Filtering strategy for expanding the search space

We discuss in Sec. 3.1 that a small space is more efficient compared to a large space for optimization given the same evaluation budget. We also discuss the problems of over- and under-specifying a search space in Sec. 3.2. Motivated by this property, we propose to start the search and expand from a given (weakly specified) space $\mathscr{X}_0$. Then, we either gradually expand the search region or exploit within the specified region.

Let us consider the maximization problem with the global maximum $f^* = \max_{\mathbf{x} \in \mathscr{X}} f(\mathbf{x})$. In our setting, however, the desired domain $\mathscr{X}$ containing the global maximum is not known in advance. Thus, during the optimization process, f^* may not be accessible. Instead, given the current space at an iteration t we only have the intermediate maximum $f_t^* = \max_{\mathbf{x} \in \mathscr{X}_t} f(\mathbf{x})$.

Since our search space over optimization iterations can only grow or stay the same, the space is non-decreasing $\mathscr{X}_0 \subseteq ... \mathscr{X}_t \subseteq \mathscr{X}_T$ and by definition $f_t^* = \max_{\mathbf{x} \in \mathscr{X}_t} f(\mathbf{x})$, we have the following lemma for the maximization problem.

Lemma 1 *The intermediate maximum is non-decreasing $f_0^* \leq ... \leq f_t^* \leq f_T^* \leq f^*$ and the intermediate gap is non-increasing $f^* - f_0^* \geq f^* - f_t^* \geq f^* - f_T^*$.*

Let x_t be our choice at an iteration t, the instantaneous regret, used in standard Bayesian optimization setting, is transformed to the case of expandable spaces as follows

$$r_t = f^* - f(\mathbf{x}_t) = \underbrace{f^* - f_t^*}_{A_t} + \underbrace{f_t^* - f(\mathbf{x}_t)}_{B_t}. \tag{6}$$

As a BO model, we aim to minimize the cumulative regret $R_T = \sum r_t$ by minimizing A_t and then B_t.

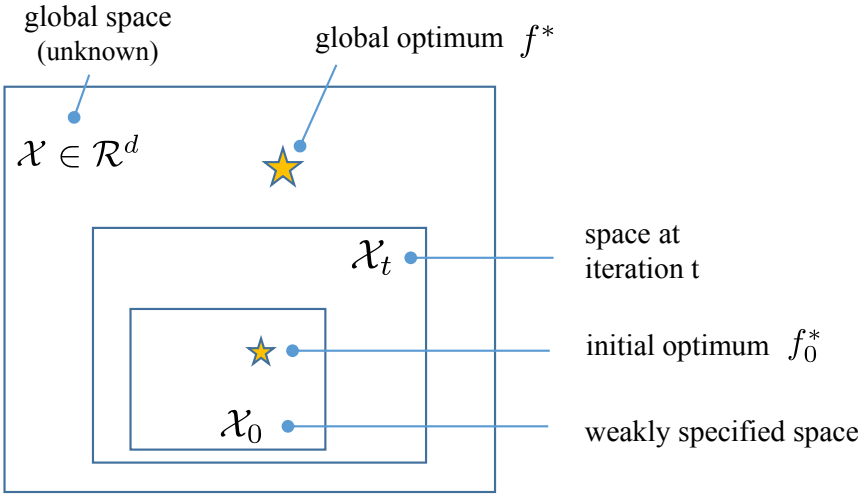


Fig. 3: Illustration of our proposed FBO. In the problem considered, the search space is unknown meaning we can not use the vanilla BO. Instead, we have the weakly specified space which brings the starting points closer to the global optimum than other arbitrary settings of the initial region.

4.1.1 Building the invasion set

We expand a space $\mathscr{X}_t$ to maximize f_t^* and thus minimize A_t. However, we also want to keep this space small (for the next step of minimizing B_t). Therefore, it is intuitive to gradually expand the search space by only selecting the promising candidates in the neighborhood, which are likely to contain better values than the current region. It is counter-intuitive to naively expand equally to all directions resulting in redundantly large search space.

Let $\mathscr{X}'_{t-1} \in \mathscr{R}^d$ be a window extending the considered space $\mathscr{X}_{t-1}$ s.t. $\mathscr{X}'_{t-1} \supset \mathscr{X}_{t-1}$. There can be different solutions in defining this neighboring window. However, any choice needs to balance the trade-off between finding better regions versus making a search space large. Specifically, we propose $\mathscr{X}'_{t-1}$ s.t. $\frac{V(\mathscr{X}'_{t-1})}{V(\mathscr{X}_{t-1})} \propto \left(\frac{h'_{t-1}}{h_{t-1}}\right)^d \triangleq \frac{T}{t-1}$ and thus a new radius is computed as $h'_{t-1} = \sqrt[d]{\frac{T}{t-1}} h_{t-1}$. This setting is reasonable to allow a greater level of exploration in the early iterations and will focus on exploitation in late iterations. If we have a big budget T, the window $\mathscr{X}'_{t-1}$ may cover the whole domain $\mathscr{X}$. In contrast, if we have a small budget, $\mathscr{X}'_{t-1}$ will slowly grow.

We select to expand a search space by discarding poor value regions. Using the GP posterior, we define the upper confidence bound (UCB) $u(\mathbf{x}) = \mu(\mathbf{x}) + \sqrt{\beta}\sigma(\mathbf{x})$ and lower confidence bound (LCB) $l(\mathbf{x}) = \mu(\mathbf{x}) - \sqrt{\beta}\sigma(\mathbf{x})$ where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the GP predictive mean and variance, respectively, in Eq. (2) and Eq. (3). From the property of GP, we have w.h.p. $l(\mathbf{x}) \leq f(\mathbf{x}) \leq u(\mathbf{x}), \forall \mathbf{x} \in \mathscr{X}_{t-1}, \mathscr{X}'_{t-1}$. The key observation is that we do not need to explore the regions (in the neighbor) that the UCB $u(\mathbf{x}')$, where $\forall \mathbf{x}' \in \mathscr{X}'_t$, is lower than the maximum value of the LCB $\max_{\forall \mathbf{x} \in \mathscr{X}_t} l(\mathbf{x})$ in the existing region. We acknowledge that the similar intuition has also been used in [10,7,3,28] for different purposes. Then, we define

the *invasion set* as the following $\mathscr{I}_t = \left\{\mathbf{x}' \in \mathscr{X}'_{t-1} \mid \mathbf{x}' \notin \mathscr{X}_{t-1}, u(\mathbf{x}') \geq \max_{\mathbf{x}\in\mathscr{X}_{t-1}} l(\mathbf{x})\right\}$ that will be used in the next step.

4.1.2 Maximizing the acquisition function

We minimize B_t in Eq. (6) by maximizing the acquisition function to select $\mathbf{x}_t$. This step is similar to the standard BO. The optimization is performed on the space of $\mathscr{X}_{t-1} \cup \mathscr{I}_t$ obtained by the above step. We find the point $\mathbf{x}_t = \underset{\mathbf{x}\in\mathscr{X}_{t-1}\cup\mathscr{I}_t}{\operatorname{argmax}}\ \alpha_t(\mathbf{x})$ maximizing the acquisition function, such as EI [25], GP-UCB [41], ES [13], PES [14], EST [43] and PVRS [29]. Note that we do not add the entire $\mathscr{I}_t$ into the new space. Instead, after obtaining $\mathbf{x}_t$, we update the new space $\mathscr{X}_t$ (as well as h_t) such that it is the smallest space containing both the old space $\mathscr{X}_{t-1}$ and the newly selected point $\mathbf{x}_t$. In other words, if $\mathbf{x}_t$ is selected in the old space $\mathscr{X}_{t-1}$, the new space will not be expanded, i.e. $\mathscr{X}_t = \mathscr{X}_{t-1}$ and $h_t = h_{t-1}$. This behavior ensures that the optimization space is not always growing for optimization efficiency with limited evaluation budget.

We summarize the proposed filtering expansion strategy for Bayesian optimization (FBO) under weakly specified space in *Algorithm* .2.

4.2 Theoretical analysis for Filtering Bayesian Optimization

The convergence rate of Bayesian optimization methods can be derived using the cumulative regret [41], the loss in reward due to not knowing f's maximum beforehand. We study the regret bound of Bayesian optimization algorithms under a weakly specified space. First, we follow [41] to provide a bound on the gap of the true function $f(\mathbf{x})$ and the GP predictive mean $\mu(\mathbf{x})$.

Lemma 2 *(Theorem 6 of [41]) Let $\delta \in (0,1)$ and define $\beta_t = 2\log\left(|\mathscr{X}|t^2\pi^2/6\delta\right)$, then*

$$P\left(\forall t, \forall \mathbf{x} \in \mathscr{X}, |\mu_t(\mathbf{x}) - f(\mathbf{x})| \leq \sqrt{\beta_t}\sigma_t(\mathbf{x})\right) \geq 1-\delta$$

Lemma 3 *(Lemma 1 of [30]) The acquisition function of EI can be expressed as $\alpha_t^{EI}(\mathbf{x}) = \sigma_{t-1}(\mathbf{x})\tau(z_{t-1}(\mathbf{x}))$ and $\alpha_t^{EI}(\mathbf{x}) \leq \tau(z_{t-1}(\mathbf{x}))$ where $\tau(z) = z\Phi(z) + \phi(z)$ with Φ and ϕ are the c.d.f. and the p.d.f. of the standard normal distribution.*

Lemma 4 *(Lemma 6 of [30]) The improvement function $I_t(\mathbf{x}) = \max\left\{0, f(\mathbf{x}) - y_{t-1}^{max}\right\}$ and the acquisition function $\alpha_t^{EI}(\mathbf{x}) = \mathbb{E}[I_t(\mathbf{x})]$ satisfy the inequality such that $I_t(\mathbf{x}) - \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}) \leq \alpha_t^{EI}(\mathbf{x})$.*

Lemma 5 *(Lemma 5.4 of [41]) Sum of the predictive variance at the selected points are bounded by the maximum information gain γ_T as $\sum_{t=1}^{T}\sigma_{t-1}^2(\mathbf{x}_t) \leq \frac{2}{\log(1+\sigma^{-2})}\gamma_T$.*

Lemma 6 *(Lemma 7 of [30]) The sum of the predictive variances is bounded by the maximum information gain γ_T. That is $\forall \mathbf{x} \in \mathscr{X}, \sum_{t=1}^{T}\sigma_{t-1}^2(\mathbf{x}) \leq \frac{2}{\log(1+\sigma^{-2})}\gamma_T$.*

Lemma 7 *(Lemma 8 of [30]) Let $\kappa > 0$ be a pre-defined stopping criteria, $z_{t-1}(\mathbf{x}) = \frac{\mu_{t-1}(\mathbf{x}) - y_{t-1}^{max}}{\sigma_{t-1}(\mathbf{x})}$ and $\tau(z) = z\Phi(z) + \phi(z)$, we have $\tau(-z_{t-1}(\mathbf{x}_t)) \leq 1 + \sqrt{C_2}$ where $C_2 \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$.*

Algorithm .2 Filtering expansion strategy for Bayesian optimization (FBO) under weakly specified space setting.

Input: #iter T, initial region $\mathscr{X}_0$ defined by a radius h_0
Randomly initialize $\mathscr{D}_0$ from $\mathscr{X}_0$.
for $t = 1$ to T **do**
  Fit a GP ϕ_t from the data $\mathscr{D}_{t-1}$.
  Build an extended window $\mathscr{X}'_{t-1}$ defined by a radius $h'_{t-1} = \sqrt[d]{\frac{T}{t-1}} \times h_{t-1}$.
  Build an invasion set $\mathscr{I}_t$ on $\mathscr{X}'_{t-1}$ given $\mathscr{X}_{t-1}$.
  Obtain $\mathbf{x}_t = \underset{\mathbf{x}\in\mathscr{X}_{t-1}\cup\mathscr{I}_t}{\operatorname{argmax}}\ \alpha_t(\mathbf{x}, \phi_t)$.
  Update the bound $\mathscr{X}_t$ and the radius h_t based on $\mathscr{X}_{t-1}$ and $\mathbf{x}_t$.
  Evaluate the function $y_t = f(\mathbf{x}_t)$.
  Augment the data $\mathscr{D}_t = \mathscr{D}_{t-1} \cup (\mathbf{x}_t, y_t)$.
end for
Output: $\mathbf{x}_{\max}, y_{\max}$

Let us denote the global optimum $f^* = \max_{\mathbf{x}\in\mathscr{X}} f(\mathbf{x})$ where $\mathscr{X}$ is assumed to be unknown or hard to specify and the initial optimum is $f_0^* = \max_{\mathbf{x}\in\mathscr{X}_0} f(\mathbf{x})$. The weakly known space condition enables that the optimum location is reachable given a limited budget. This assumption is essential to have a rigorous bound on the cumulative regret. If the initial region is poorly placed, it may get stuck at low-value regions and thus the regret bound may not be bounded properly.

We use ω_t to denote the gap between the global optimal and the optimal in the intermediate region as $f^* - f_t^* = \omega_t$. We below derive the regret bound for the Bayesian optimization algorithms (including vanilla BO and our FBO) under the weakly specified space.

Theorem 1 *Let* $\beta_T = 2||f||_k^2 + 300\gamma_t \ln^3\left(\frac{T}{\delta}\right)$, $\delta \in (0,1)$, $C_1 = \frac{2\beta_T}{\log\left(1+\sigma^{-2}\right)}$, $C_2 \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$, κ *be the stopping threshold used in EI (to make the convergence proof feasible) and the maximum information gain* γ_T. *Then with probability at least* $1-\delta$, *the cumulative regret of the Bayesian optimization algorithms under the weakly specified setting are bounded for GP-UCB as* $R_T^{UCB} \leq \sqrt{C_1 T \gamma_T} + \sum_{t=1}^{T} \omega_t$ *and for EI as* $R_T^{EI} \leq \sqrt{\frac{2T\gamma_T}{\log\left(1+\sigma^{-2}\right)}}\left[\sqrt{3\left(\beta_T + 1 + C_2\right)} + \sqrt{\beta_T}\right] + G_T$.

Proof Since the derivation for two acquisition functions GP-UCB and EI are different, we first present the proof for GP-UCB, then we derive for EI.

i) For the first case of GP-UCB, we start with the instantaneous regret using GP-UCB as the acquisition function

$$\begin{aligned} r_t &= f^* - f(\mathbf{x}_t) = f_t^* - f(\mathbf{x}_t) + \omega_t \\ &= f_t^* - \mu_{t-1}(\mathbf{x}^*) + \mu_{t-1}(\mathbf{x}^*) - f(\mathbf{x}_t) + \omega_t \\ &\leq \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}^*) + \mu_{t-1}(\mathbf{x}^*) - f(\mathbf{x}_t) + \omega_t \quad \text{by Lem 2} \\ &\leq 2\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}_t) + \omega_t \end{aligned}$$

where the last term is obtained by using that $\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}^*) + \mu_{t-1}(\mathbf{x}^*) \leq \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}_t) + \mu_{t-1}(\mathbf{x}_t)$ since $\mathbf{x}_t = \underset{\mathbf{x}\in\mathrm{X}}{\operatorname{argmax}}\alpha^{\mathrm{UCB}}(\mathbf{x})$ and then utilize Lem. 2 again. By using Lem. 5, we

bound the sum of variances using the maximum information gain and show that

$$\sum_{t=1}^{T} 4\beta_t \sigma_{t-1}^2(\mathbf{x}_t) \leq \frac{2 \times \beta_T \gamma_T}{\log(1+\sigma^{-2})}.$$

Next, we utilize the Cauchy-Schwartz to have

$$\sum_{t=1}^{T} 2\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}_t) \leq \sqrt{\frac{2 \times T \times \beta_T \times \gamma_T}{\log(1+\sigma^{-2})}}. \tag{7}$$

We have the cumulative regret bound $R_T = \sum_t^T r_t$ under the weakly specified space setting as

$$\begin{aligned} R_T &\leq \sum_{t=1}^{T} 2\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}_t) + \sum_{t=1}^{T} \omega_t \\ &\leq \sqrt{C_1 T \gamma_T} + G_T \qquad \text{by Eq. (7)} \end{aligned} \tag{8}$$

where $C_1 = \frac{2\beta_T}{\log(1+\sigma^{-2})}$ and $G_T = \sum_{t=1}^T \omega_t$.

ii) For the second case of EI, we begin with the instantaneous regret as

$$\begin{aligned} r_t &= f(\mathbf{x}^*) - f(\mathbf{x}_t) = f_t^* - f(\mathbf{x}_t) + \omega_t \\ &= \underbrace{f_t^* - y_{t-1}^{\max}}_{A_t} + \underbrace{y_{t-1}^{\max} - f(\mathbf{x}_t)}_{B_t} + \omega_t. \end{aligned} \tag{9}$$

We bound r_t with the GP posterior variance so that we later connect it to the maximum information gain γ_T. From the definition of $\mathbf{x}_t$ and Lem. 3, we have $\alpha_t^{\text{EI}}(\mathbf{x}^*) \leq \alpha_t^{\text{EI}}(\mathbf{x}_t) = \sigma_{t-1}(\mathbf{x}_t)\,\tau(z_{t-1}(\mathbf{x}_t))$. Then, by using Lem. 4 we write

$$\begin{aligned} A_t &\leq \alpha^{\text{EI}}(\mathbf{x}^*) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}^*) \leq \alpha^{\text{EI}}(\mathbf{x}_t) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}^*) \\ &= \sigma_{t-1}(\mathbf{x}_t)\,\tau(z_{t-1}(\mathbf{x}_t)) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}^*) \qquad \text{by Lem. 2} \end{aligned}$$

Next, we express the second term in Eq. (9) as follows

$$\begin{aligned} B_t =& y_{t-1}^{\max} - \mu_{t-1}(\mathbf{x}_t) + \mu_{t-1}(\mathbf{x}_t) - f(\mathbf{x}_t) \\ \leq& \sigma_{t-1}(\mathbf{x}_t)\,(-z_{t-1}(\mathbf{x}_t)) + \sigma_{t-1}(\mathbf{x}_t)\sqrt{\beta_t} \qquad \text{by Lem. 2} \\ =& \sigma_{t-1}(\mathbf{x}_t)\left[\tau(-z_{t-1}(\mathbf{x}_t)) + \sqrt{\beta_t} - \tau(z_{t-1}(\mathbf{x}_t))\right] \qquad \text{by } z = \tau(z) - \tau(-z) \end{aligned}$$

Continuing from Eq. (9), we have $r_t = A_t + B_t$ that is

$$r_t \leq \sigma_{t-1}(\mathbf{x}_t)\left[\sqrt{\beta_t} + \tau(-z_{t-1}(\mathbf{x}_t))\right] + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}^*) \tag{10}$$

Using the bound of $\tau(-z_{t-1}(\mathbf{x}_t))$ in Lem. 7, we obtain

$$r_t \leq \underbrace{\sigma_{t-1}(\mathbf{x}_t)\left[\sqrt{\beta_t} + 1 + \sqrt{C}\right]}_{L_t} + \underbrace{\sqrt{\beta_t}\sigma_{t-1}(\mathbf{x}^*)}_{U_t}$$

where $C_2 \triangleq \log\left[\frac{1}{2\pi\kappa^2}\right]$. We then simplify L_t and U_t, respectively. Taking the sum of squared regret and utilizing the Cauchy-Schwartz inequality that $(a+b+c)^2 \leq 3\left(a^2+b^2+c^2\right)$, we have

$$\begin{aligned}
\sum_{t=1}^{T} L_t^2 &\leq \sum_{t}^{T} \sigma_{t-1}^2(\mathbf{x}_t)\, 3\left(\beta_t + 1 + C_2\right) \\
&\leq 3\left(\beta_T + 1 + C_2\right) \sum_{t}^{T} \sigma_{t-1}^2(\mathbf{x}_t) && \text{by } \beta_T \geq \beta_t, \forall t \leq T \\
&\leq \frac{6\left(\beta_T + 1 + C_2\right)\gamma_T}{\log\left(1+\sigma^{-2}\right)} && \text{by Lem. 6}
\end{aligned}$$

Again, using the Cauchy-Schwartz inequality, we obtain

$$\sum_{t=1}^{T} L_t \leq \sqrt{T}\sqrt{\sum_{t=1}^{T} L_t^2} \leq \sqrt{\frac{6T\left(\beta_T + 1 + C_2\right)\gamma_T}{\log\left(1+\sigma^{-2}\right)}}.$$

We further utilize Lem. 6 and the Cauchy-Schwartz again to simplify U_t that

$$\sum_{t=1}^{T} U_t \leq \beta_T \sum_{t=1}^{T} \sigma_{t-1}(\mathbf{x}^*) \leq \sqrt{\frac{2T\beta_T\gamma_T}{\log\left(1+\sigma^{-2}\right)}}.$$

Finally, we get the cumulative regret $R_T \leq \sum_{t=1}^{T}\left(L_t + U_t\right) + \sum_{t=1}^{T}\omega_t$,

$$\begin{aligned}
R_T &\leq \sqrt{\frac{2T\gamma_T}{\log\left(1+\sigma^{-2}\right)}}\left[\sqrt{3\left(\beta_T + 1 + C_2\right)} + \sqrt{\beta_T}\right] + \sum_{t=1}^{T}\omega_t \\
&\leq \sqrt{\frac{2T\gamma_T}{\log\left(1+\sigma^{-2}\right)}}\left[\sqrt{3\left(\beta_T + 1 + C_2\right)} + \sqrt{\beta_T}\right] + G_T \qquad (11)
\end{aligned}$$

where $G_T = \sum_{t=1}^{T}\omega_t$.

The derived regret bound in Theorem 1 is general for multiple BO algorithms considering the situation that the global optimum may not include in the predefined space. The regrets in Eq. (8) and Eq. (11) are bounded by a sublinear term and a sum of constants ω_t which depends on how good the initial region is located and the expanding behavior. Ideally if the initial space contains the global optimum (or $\omega_0 = 0$), we will achieve the sublinear rate $R_T \leq \sqrt{CT\gamma_T}$ as in the previous work [41]. However, we note that Theorem 1 on its own does not guarantee the sublinear rate of the algorithm under this weakly specified space, i.e. $\lim_{T\to\infty}\frac{R_T}{T} \neq 0$.

We now discuss the implication of the regret bound for vanilla BO and our FBO from Eq. (8) and Eq. (11). The regret bounds of BO and FBO are different from two factors: G_T and γ_T.

In terms of G_T, we have $G_T^{\text{BO}} = \omega_0 T$ for vanilla BO because of the fixed space $\mathscr{X}_0$. In contrast, since our approach can expand the space, we have $\omega_0 \geq \omega_t \geq \omega_T$ (using Lem. 1) and thus $\sum_{t=1}^{T}\omega_t \leq \omega_0 T$. As a result, $G_T^{\text{FBO}} \leq G_T^{\text{BO}}$. This is an advantage of FBO against the vanilla BO.

In terms of γ_T, we have in Eq. (4) that $\gamma_T \leq \left[\log\left(Th\right)^d T\right]^{d+1}$ where h is a radius of the space. Our approach allows the search space to grow while the vanilla BO does not. Thus, we have $h^{\text{FBO}} \geq h^{\text{BO}}$ and $\gamma_T^{\text{FBO}} \geq \gamma_T^{\text{BO}}$.

We note that other expanding schemes in Bayesian optimization, including Volume doubling, also share a similar form of regret bounds in Eq. (8) and Eq. (11). However, the search space in Volume doubling is often larger than required. This becomes inefficient for optimization, especially when the evaluation budget is always limited.

4.3 Optimizing Gaussian process hyper-parameter

For robustness, we estimate the GP hyper-parameters by maximizing their posterior probability (MAP) as the marginal likelihood, $p(\theta \mid \mathbf{X}, \mathbf{y}) \propto p(\theta, \mathbf{X}, \mathbf{y})$, which, thanks to the Gaussian likelihood, is available in closed form as [34]

$$\ln p\left(y, X, \theta\right) = -\frac{1}{2}\ln\left|K+\sigma^2 I_n\right| - \frac{1}{2}y^T\left(K+\sigma^2 I_n\right)^{-1} y + \ln p_0\left(\theta\right) + const$$

where I_n is the identity matrix in dimension n (the number of points in the training set) and $p_0\left(\theta\right)$ is the prior over hyper-parameters.

5 Experiments

We first illustrate the expansion and optimization behavior of the proposed FBO on 2D function to gain insight understanding. Next, we evaluate our method on 8 benchmark functions and two real-world applications: machine learning hyper-parameter tuning and experimental alloy design. To gain further insight, we discuss the cumulative regret and computational complexity of different approaches. All the **source codes** are available for reproducibility at the link https://github.com/ntienvu/ICDM2017_FBO.

5.1 Experimental setting

Given dimension d, the optimization is run with an evaluation budget of $T = 10d$ excluding an initial $3d$ points (as used in [31]). We repeat the experiments 20 times and report the mean and standard error. We use the squared exponential kernel $k\left(x, x'\right) = \exp\left(-\frac{||x-x'||^2}{\sigma_l^2}\right)$ where σ_l^2 is optimized using maximizing the marginal likelihood (see Sec. 4.3), the input x is normalized as $[0,1]$ and the output is standardized $y \sim \mathcal{N}\left(0,1\right)$ for robustness. We always maximize the objective function, maximizing $-f$ for cases in which the goal is to find the minimum. For methods using the UCB, we set $\sqrt{\beta} = 2$ (as also used in GPyOpt [11] and [31]), which allows us to compare the different batch methods using the same acquisition function. All implementations are in Python. All simulations are done on Windows machine Core i7 Ram 24GB.

5.2 Visualization of filtering strategy for BO

Before presenting the numerical results, we visualize the expansion behavior of FBO in Fig. 5 using a Branin function as the objective function. These illustrations are particularly steps 4-7 of Algorithm .2. In Fig. 5 (top), we plot the 3D view with the initial region $\mathscr{X}_0$. Then, we visualize the old space $\mathscr{X}_{t-1}$ and the invasion set $\mathscr{I}_t$ consisting of the promising points

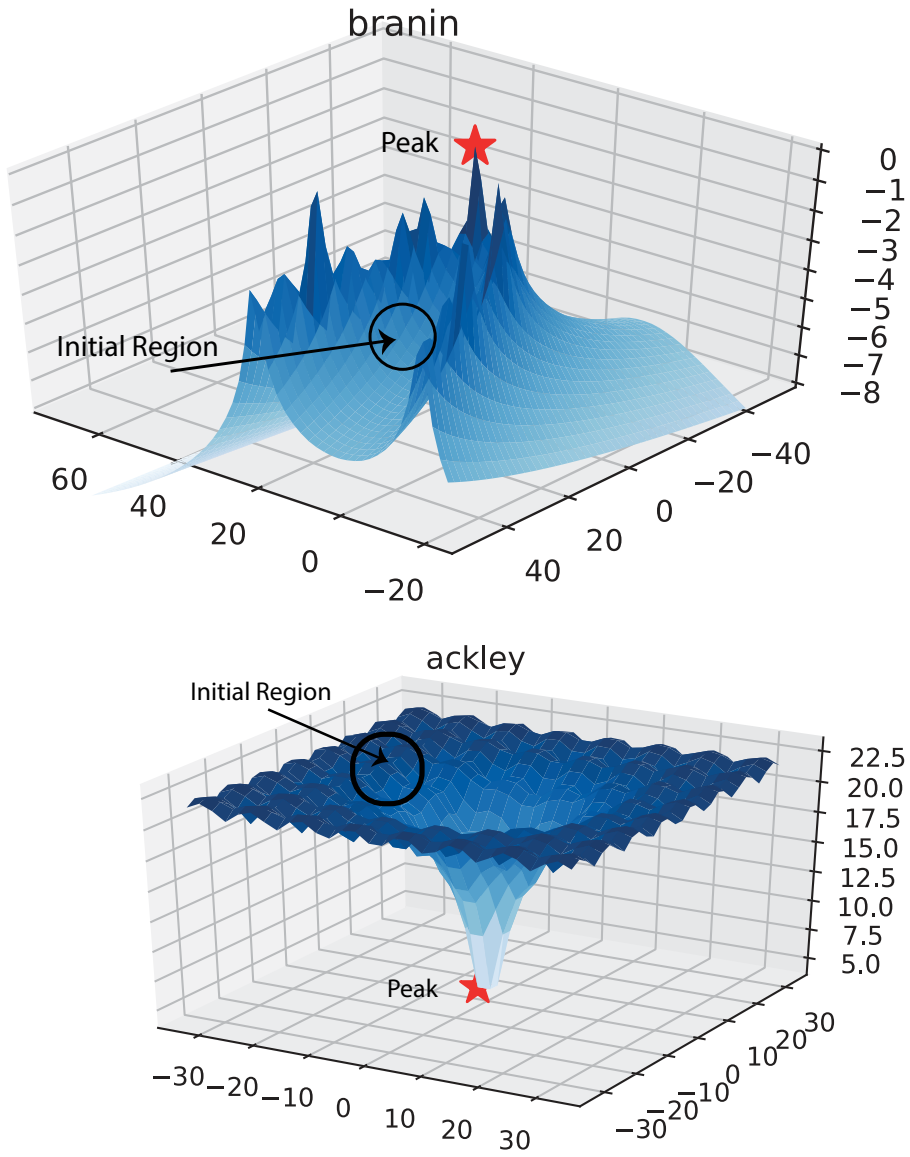


Fig. 4: Two examples of Branin function and Ackley function. The red stars are the peak locations. The initial regions do not contain the peak.

(in magenta) (left) at different iterations. We illustrate the new space $\mathscr{X}_t$ which expands the previous one (right). As illustrated, instead of expanding naively in all directions, the space is selected to extend toward high-value regions the bottom right for $t = 12$ and toward the top for $t = 16$.

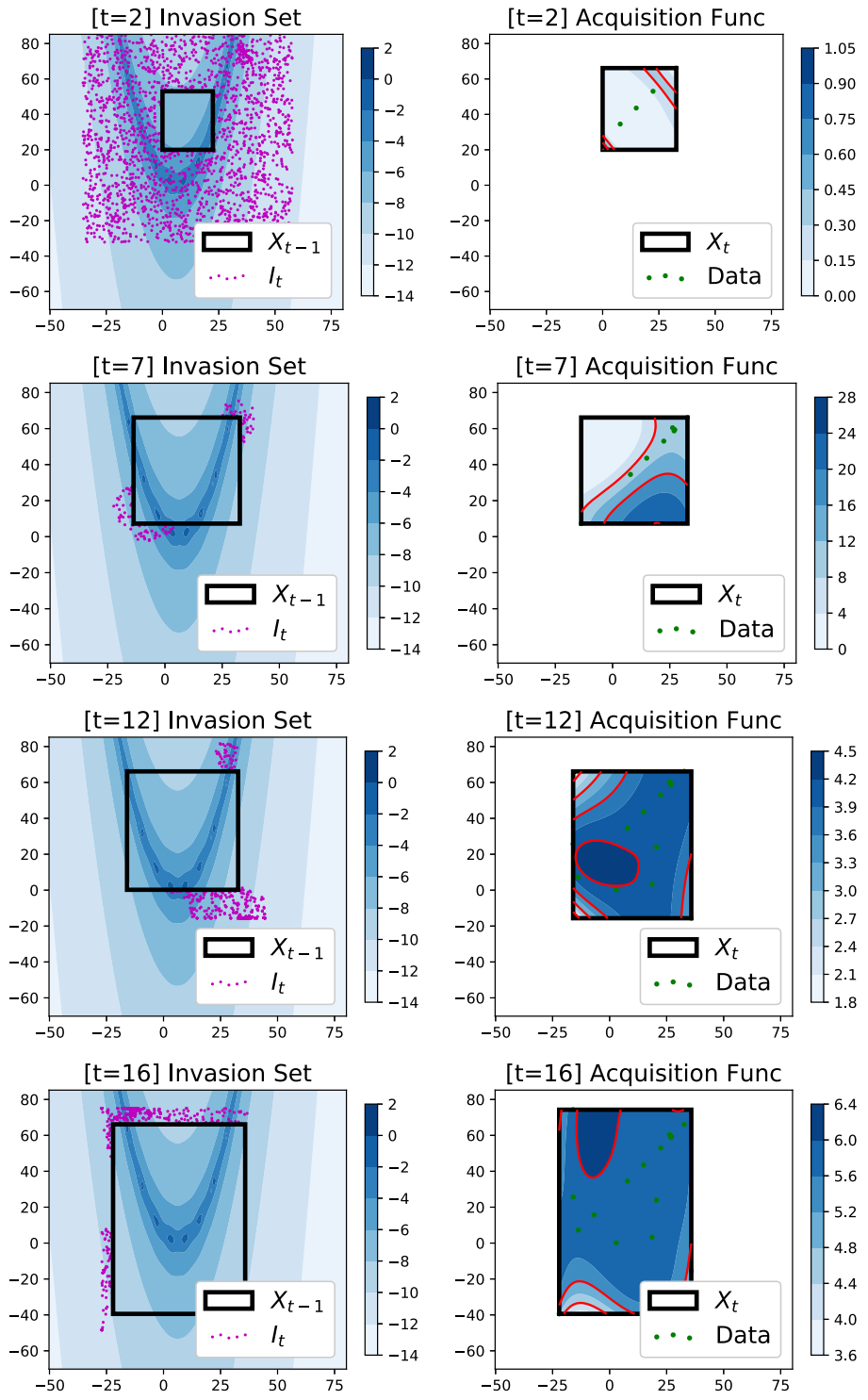


Fig. 5: FBO visualization using Branin 2D function. Left: the original function $f(x)$ overlaid with the space $\mathscr{X}_{t-1}$ and an invasion set $\mathscr{I}_t$. Each magenta dot is a point in $\mathscr{I}_t$. Right: Acquisition function $\alpha(x)$ with a new point added within the expanded space $\mathscr{X}_t$ and the observations $\mathscr{D}_t$. (best viewed in color).

Table 2: FBO achieves the best performances in 6 out of 8 cases. The initial space $\mathscr{X}_0$ is created by taking $[10\%, 30\%]$ of the original spaces - available for each function. We denote the minimum (for minimization problem) in the initial region as $f_0^* = \min_{\forall \mathbf{x} \in \mathscr{X}_0 \subset \mathscr{X}} f(\mathbf{x})$. Vanilla BO performs optimization restrictively on $\mathscr{X}_0$ while the others will start from $\mathscr{X}_0$ and expand the space. *M* indicares million unit for gSobol function.

Function	Func	Beale	Six-humpcamel	Rosenbrock	Branin
	Dim	2	2	2	2
	f_0^*	24.63	1.33	21.23	21
Vanilla BO	UCB	***26.8±0***	1.331±0.0	21.2±0	72.1±0
	EI	25.5±200	1.331±0.0	21.2±.1	72.1±0
Volume Doubling	UCB	**20.4±32**	0.42±0.8	5.62±7	7.16±5
	EI	48±91	11.25±2.4	20±42	5.43±5
Regularize BO	EI-H	85±190	0.101±.86	94±115	8.97±8
	EI-Q	264±1k	1.25±2.6	93±84	10.1±7
FBO	UCB	40±59	**-0.296±.21**	***5.48±5***	**2.57±2**
	EI	36.4±56	-0.109±.37	**3.09±3**	***2.88±3***

Function	Func	Hartmann	Ackley	Hartmann	gSobol
	Dim	3	5	6	10
	f_0^*	-0.986	19.08	1.89	0
Vanilla BO	UCB	-.96±.1	19.6±.3	-1.88±0	.6M±.1M
	EI	-.97±.01	19.4±.3	-1.88±.1	.6M±.1M
Volume Doubling	UCB	-2.69±.7	20.1±.6	-2.67±.1	.6M±.2M
	EI	-2.54±.8	19.8±1	-2.75±.1	.7M±.1M
Regularize BO	EI-H	-1.95±.9	19.6±1	-2.53±.1	3.8M±1M
	EI-Q	-2.13±.8	20.0±.2	-2.54±.1	3.5M±3M
FBO	UCB	***-2.91±.6***	***17.1±2***	***-2.79±.1***	**.5M±.2M**
	EI	**-3.24±.4**	**17.2±3**	**-2.88±.1**	**.5M±.2M**

5.3 Baselines

We create the initial region $\mathscr{X}_0$ as follows. Let $\mathscr{X}$ be the pre-defined search space for each function (e.g., $[0,1]^6$ for Hartmann function). We define the initial region $\mathscr{X}_0$ as $[10\% - 30\%]$ of the pre-defined space (e.g., $[0.1, 0.3]^6$ for Hartmann function). We utilize the following baselines for comparison.

- *Vanilla BO:* Bayesian optimization is restricted on the initial space $\mathscr{X}_0$.
- *Volume doubling:* We naively double the volume after $2d$ evaluations starting from the initial region $\mathscr{X}_0$. This approach is also used in [36] as a baseline for comparison.
- *Regularizing approaches* [36]*:* EI-H: Regularized EI with a hinge-quadratic prior, we set $\beta = 1$ and R fixed by the circumradius of the initial box, as used in [36]. EI-Q: Regularized EI with a quadratic prior mean where the widths w are fixed to those of the initial bounding box as in [36]. The optimization starts from the initial region $\mathscr{X}_0$.

5.4 Evaluation on benchmark functions

Our first set of experiments is validating FBO on benchmark objective functions. Given the weakly specified space settings, we demonstrate that FBO outperforms all of the baselines.

We select 8 popular benchmark functions. Details of these functions are available at the link[2].

We present the numerical results in Table 2 across dimensions (2 to 10). Vanilla BO settings using UCB and EI perform poorly in this setting. The reason is that vanilla BO only performs optimization on the fixed region $\mathscr{X}_0$ which may not contain the optimum. We note that if the optimum is ideally located in the initial box, the vanilla BO using a small space will achieve the best performance.

Naive volume doubling is simple to implement. However, volume doubling is inefficient for Bayesian optimization as it does not take into account the knowledge of the optimization process to select promising directions to expand. Instead, it expands equally in all directions, thus making a search space larger than required. Moreover, setting the parameter corresponding to the number of iterations to expand volume is critical and hard to specify.

Although regularizing algorithms [36] can select points outside the initial region, it is sensitive to a box center, $\bar{x}$. Thus, they are less prone to expand toward high-value regions. In addition, EI-H and EI-Q require additional parameters (R, β and w) that are sensitive and difficult to specify in practice. Moreover, there is no theoretical guarantee for these regularizing approaches.

The results confirm our hypothesis that the proposed strategy is capable of useful exploration outside the initial region. Our FBO utilizes the property of GP to expand towards higher value regions. This strategy flexibly allows our algorithm to stop expanding and focus exploiting when there is no promising region available. In contrast, the volume doubling approach will keep expanding continuously although the evaluation budget is always limited. Thus, FBO is more efficient for optimization, utilizing the pre-defined space and expanding this space towards the promising regions. We also present the performance w.r.t. iterations for the Sixhump camel 2D, Ackley 5D and Hartmann 6D in Fig. 6.

5.5 Average regret analysis

To gain insight, we further study the average regret of our model and the considered baselines in Fig. 7 using Sixhump camel 2D, Ackley 5D and Hartmann 6D. In particular, we learn that the regularized approaches (EI-Q, EI-H) and volume doubling fluctuate and tend to increase the cumulative regret due to the exploration effect in their expansion. The vanilla BO performs optimization only on the initial space, which may contain relatively good values. Thus, the average regret of the vanilla BO is more stable and better than the regularized and volume doubling. Our proposed filtering approach selects to expand toward high-value regions and may stop the expansion if there is no promising region available. Thus, FBO achieves better average regret than the other methods (cf. Fig. 7).

5.6 Computational time comparison

We study the computational time spent per iteration w.r.t. increasing dimensions from 5 to 10. Comparing to the vanilla BO, FBO takes an extra step for building the invasion set (step 4 in Algorithm .2) to expand the space. This step makes FBO slower than the vanilla BO. However, it should be noted in the context of Bayesian optimization that the time for evaluating the black-box function (e.g., evaluating a real alloy testing) is much more expensive

[2] `https://www.sfu.ca/~ssurjano`

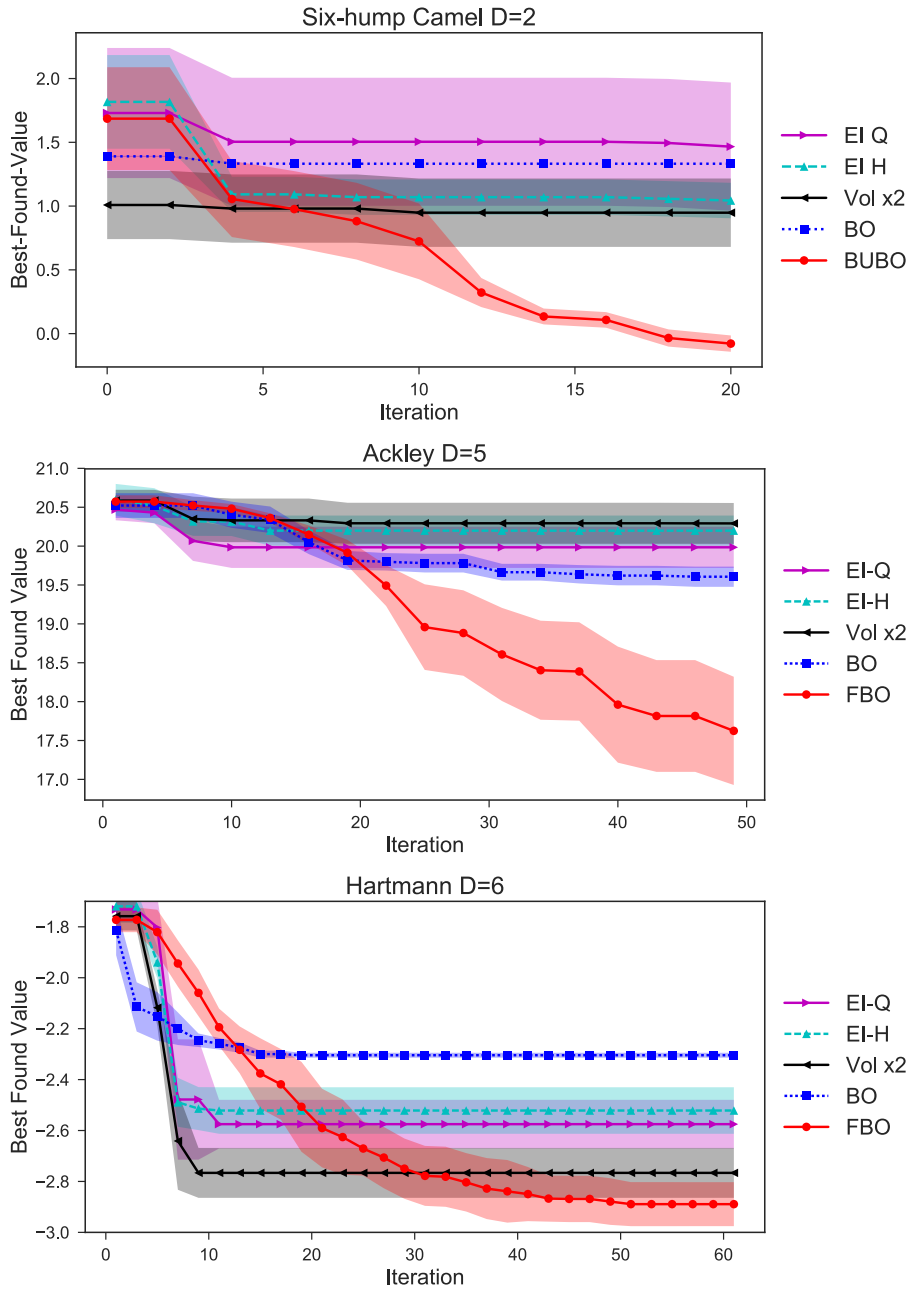


Fig. 6: Performance comparison w.r.t. iterations using best found value $\min_{i=1\ldots t-1} f(x_i)$. Our approach outperforms the baselines in finding the optimum for Six-hump camel 2D, Ackley 5D and Hartmann 6D functions. The superior performance of FBO is the results of expanding the search space toward the promising location.

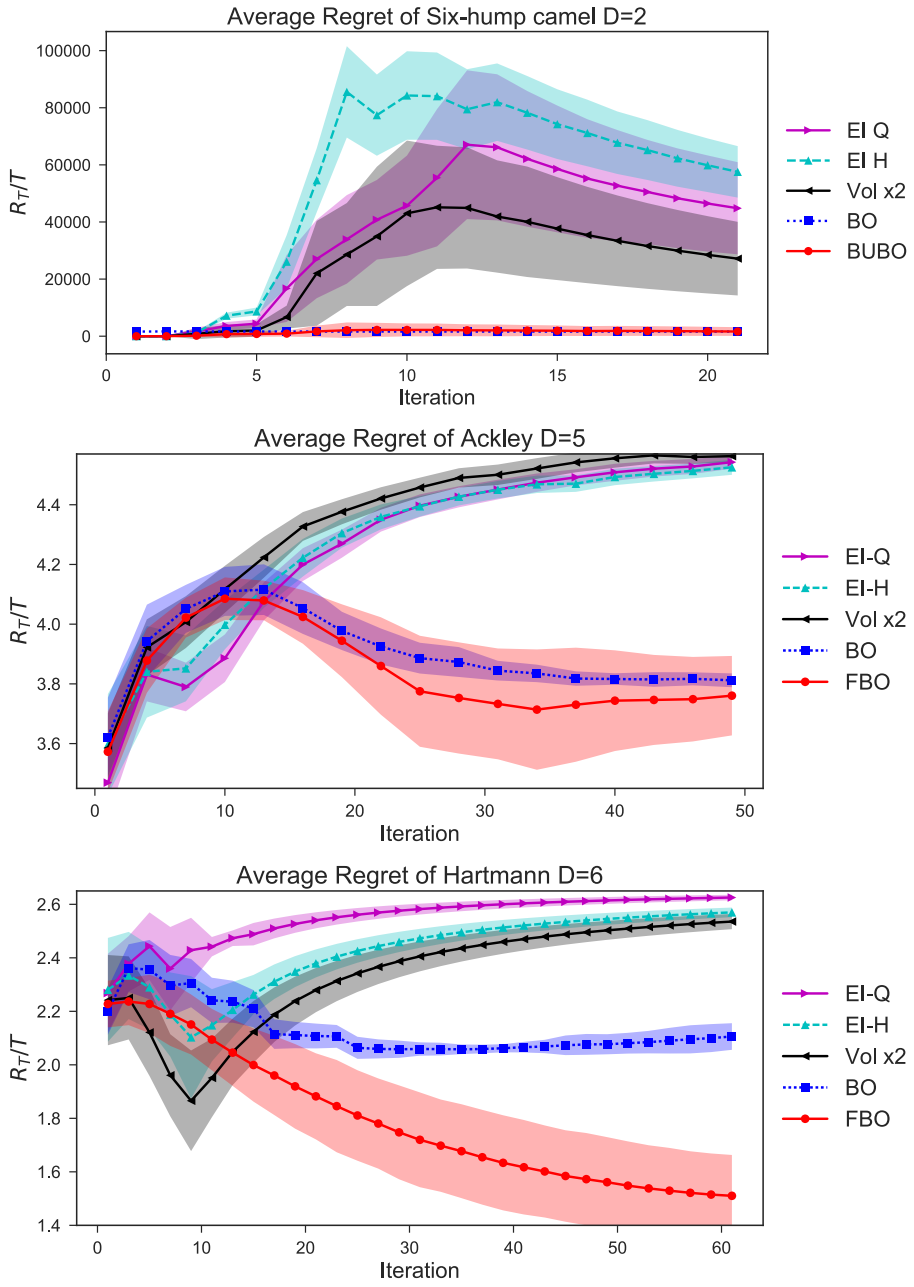


Fig. 7: The average regret is defined as $\frac{R_T}{T}$ where $R_T = \sum_{t=1}^{T} r_t$. The regularized approaches and volume doubling fluctuate and tend to increase the average regret due to their exploration effects in their expansion. The vanilla BO performs optimization only on the initial space, thus BO's performance is restricted. Our FBO selects to expand toward high-value regions and may stop expanding if there is no promising region available. Therefore, FBO obtains better average regret than the others.

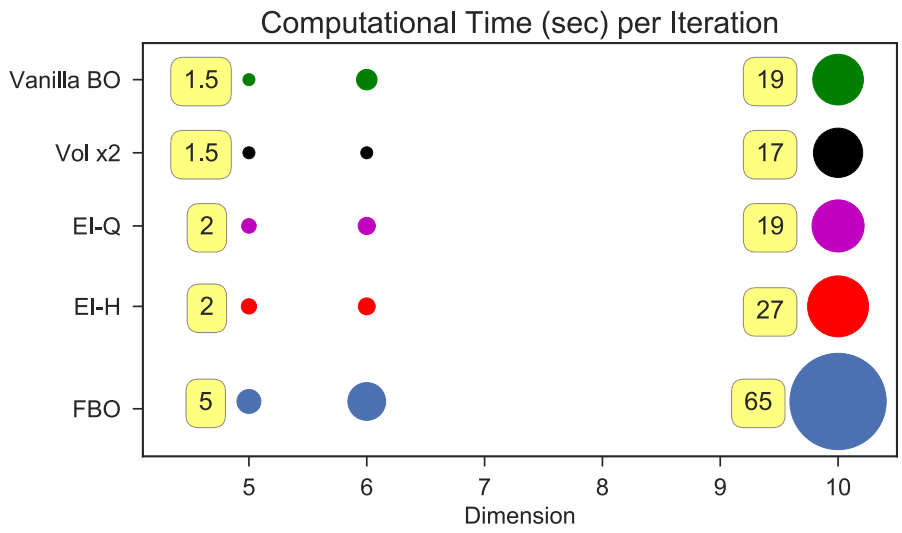


Fig. 8: Computational time per iteration. The circle indicates the magnitude. The exact times are annotated for dimensions 5 and 10. Although FBO is taking more computational time, it gains efficiency in finding the optimum for unknown space setting. In the context of optimizing the expensive black-box functions, the CPU time of FBO is negligible.

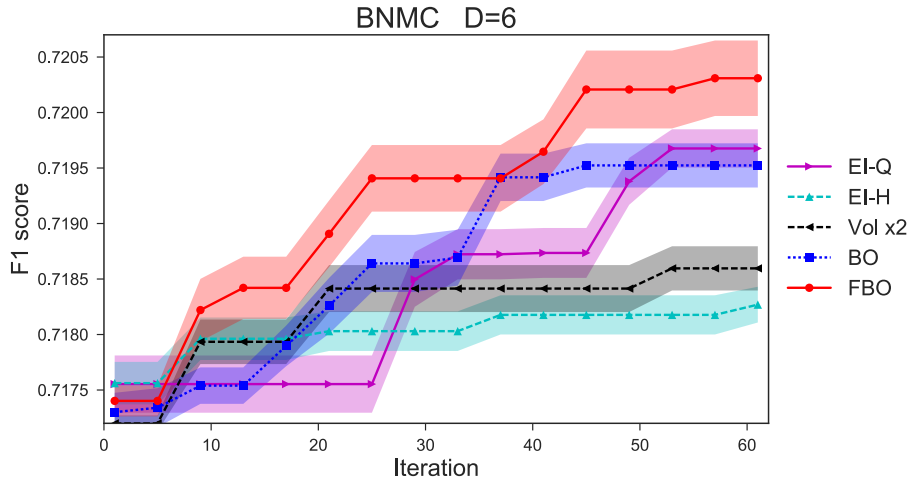
than the time for optimizing BO. Thus, our approach is still efficient for BO, especially for optimizing the expensive and time consuming real-world experiments, although there is an extra CPU cost. We plot the wall-clock time per iteration (in seconds) in Fig. 8 - the CPU time for FBO is generally 3-4 times higher than the baselines.

5.7 Machine learning hyper-parameter tuning

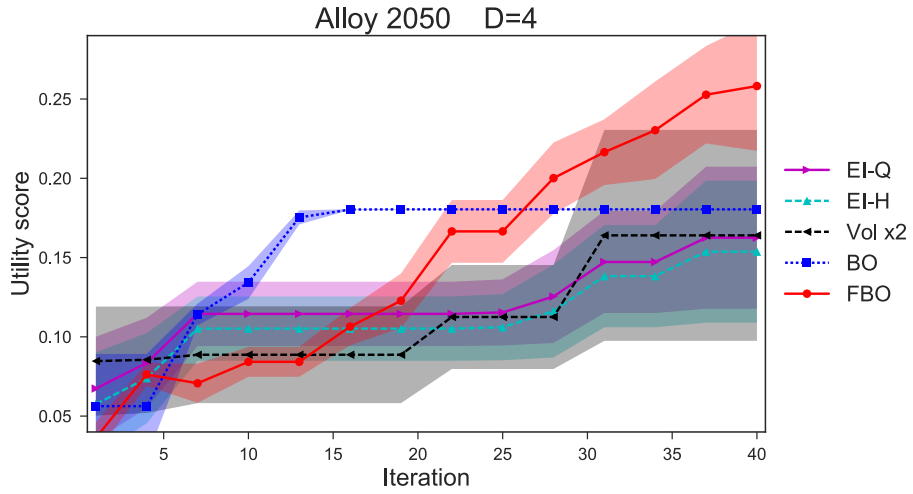
We tune the hyper-parameters for multi-label classification machine learning algorithm of Bayesian nonparametric multi-label classification (BNMC) [26] on the Scene dataset using the released source code, constructed as a black-box function[3]. BNMC uses stochastic variational inference (SVI) and stochastic gradient descent (SGD) for learning. In particular, we optimize 6 hyper-parameters for BNMC: Dirichlet symmetric for features and labels, learning rate for SVI and SGD, truncation threshold and stick-breaking parameters. We aim to maximize the F1-score.

The spaces which contain the optimal value for BNMC are unknown, instead we have the default setting [26] that returns the F1-score of 0.705. Although this value is not yet optimal, it is relatively accurate (w.r.t. other baselines in multi-label classification literature). Thus, it makes sense to define an initial region $\mathscr{X}_0$ from this default setting and consider it as a weakly specified space. Then, we let FBO expand the search and perform optimization to obtain the best F1-score of 0.72. We present the performance w.r.t iterations in Fig. 9a.

[3] https://github.com/ntienvu/ACML2016_BNMC



(a) Machine learning hyper-parameter tuning.



(b) Designing alloy 2050 for aeronautical industry.

Fig. 9: Performance comparison on real-world applications w.r.t. iterations. The higher F1 score and utility score indicate the better algorithm. The initial region of BNMC is defined using the default parameters used in [26]. The initial region of alloy 2050 is weakly defined by the metallurgist collaborators. These results are using EI as the acquisition function while the UCB results are omitted to avoid the clutter in the graph.

5.8 Experimental design

We consider the low density alloy AA-2050 [22, 1] used in the aeronautical industry. This alloy offers a low density high corrosion resistant alternative to incumbent medium to thick plate alloys and to thin plate alloys. The considered alloy consists of 8 elements (Al, Cu, Li,

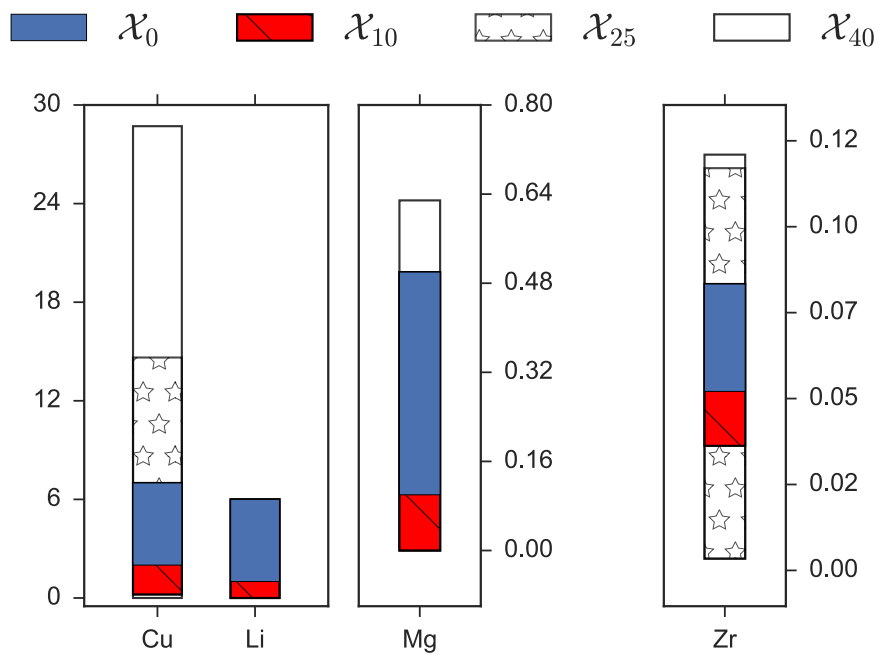


Fig. 10: The space expansion behavior in alloy design. We record the search space $\mathscr{X}_t$ at iterations $\{0, 10, 25, 40\}$ for four elements {Cu,Li,Mg,Zr} that the initial region $\mathscr{X}_0$ is provided by our metallurgist collaborators. The y-axis indicates the upper and lower ranges of the space (in percent) for each element in the AA-2050 composition w.r.t. considered iterations. (best view in color).

Mg, Zr, Sc, Si and Fe) while the Sc, Si and Fe are fixed values in our problem that we do not need to optimize. We have the constraint that the sum of all elements is 100%. Thus, we select to optimize four elements (Cu,Li,Mg,Zr) and treat the remaining element of Al as the dependent variable for simplicity.

We aim to find the AA-2050 composition to achieve the desired properties, such as low-density high corrosion resistant. The desired property for the alloy is defined using the utility score which includes maximizing good phases while minimizing the bad phases at equilibrium of a heat treatment process [20]. The good and bad phases are designed by our metallurgist collaborators. We measure these phases using the software of Thermocalc[4] at different temperatures (160, 300, 420 and 500 degrees Celsius). We then summarize these outputs and compute the utility score y.

We aim to optimize the utility score by finding the best alloy composition. However, the search space containing the optimal value for the chosen elements is unknown and hard to specify due to limited knowledge. Our metallurgist collaborators can only suggest a vague space as the initial bounds for the optimization.

We present the quantitative comparison in Fig. 9b. Vanilla BO obtains the smallest variance due to performing optimization within a small space. Our approach outperforms the

[4] `http://www.thermocalc.com`

others by a wide margin in the utility achieved after iteration 25. To gain understanding, we plot the space expansion behavior by FBO in Fig. 10. We notice that FBO adds more Cu into the composition (from 14% to 30%) after 25 iterations while it prevents the expansion of Li, Mg, and Zr, ensuring small, but efficient, search space. In addition, FBO only slightly expands Li space downwards within the first 10 iterations, then stops the expansion and focuses on exploitation. This phenomenon is the result of our expansion strategy with the invasion set that smaller Li is not the promising direction to explore.

6 Discussion

We consider a possible situation when our FBO can be failed due to poor initialization or the nature of the function. This case can happen when the large valley is surrounding the current space which can prevent the expansion of FBO if we have the limited evaluation budget T. For addressing this issue, in a future work section, we will investigate alternative expansion strategies so that we can get rid of this valley trap.

7 Conclusion

We have presented a new strategy for Bayesian optimization in weakly specified search space. Our approach can be applied when the search space is not well defined. Indeed, given an initial region that does not include the optimum, we have demonstrated that our approach can expand its region of interest and achieve greater function values. Our method contributes toward the current Bayesian optimization framework for many practical applications, and can be readily used with any acquisition function which is induced by a GP.

8 Future Work

We suggest four additional interesting directions to pursue. First, we will investigate alternative expansion strategies. Second, we consider the advanced setting where the initial region is not weakly specified, but placed arbitrarily. Third, we will devise the batch Bayesian optimization using filtering strategy for the settings that parallel evaluations are possible. Fourth, we will estimate and utilize the partial derivative information to move, expand or shrink the search space toward the global optimum location.

9 Acknowledgments

This research was partially funded by the Australian Government through the Australian Research Council (ARC) and the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning. Prof Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

References

1. M-N Avettand-Fènoël and R Taillard. Effect of a pre or postweld heat treatment on microstructure and mechanical properties of an aa2050 weld obtained by ssfsw. *Materials & Design*, 89:348–361, 2016.

2. Prasanna V Balachandran, Dezhen Xue, James Theiler, John Hogden, and Turab Lookman. Adaptive strategies for materials design using uncertainties. *Scientific reports*, 6, 2016.
3. Ilija Bogunovic, Jonathan Scarlett, Andreas Krause, and Volkan Cevher. Truncated variance reduction: A unified approach to bayesian optimization and level-set estimation. In *Advances in Neural Information Processing Systems*, pages 1507–1515, 2016.
4. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
5. Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.
6. Adam D Bull. Convergence rates of efficient global optimization algorithms. *The Journal of Machine Learning Research*, 12:2879–2904, 2011.
7. Emile Contal, David Buffoni, Alexandre Robicquet, and Nicolas Vayatis. Parallel gaussian process optimization with upper confidence bound and pure exploration. In *Machine Learning and Knowledge Discovery in Databases*, pages 225–240. Springer, 2013.
8. Thanh Dai Nguyen, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, Kyle J Deane, and Paul G Sanders. Cascade Bayesian optimization. In *Australasian Joint Conference on Artificial Intelligence*, pages 268–280. Springer, 2016.
9. Peter I Frazier and Jialei Wang. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pages 45–75. Springer, 2016.
10. Nando D Freitas, Masrour Zoghi, and Alex J Smola. Exponential regret bounds for gaussian process bandits with deterministic observations. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1743–1750, 2012.
11. Javier González, Zhenwen Dai, Philipp Hennig, and Neil D Lawrence. Batch Bayesian optimization via local penalization. In *International Conference on Artificial Intelligence and Statistics*, pages 648–657, 2016.
12. Javier Gonzalez, Michael Osborne, and Neil Lawrence. Glasses: Relieving the myopia of Bayesian optimisation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 790–799, 2016.
13. Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13:1809–1837, 2012.
14. José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pages 918–926, 2014.
15. José Miguel Hernández-Lobato, James Requeima, Edward O Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. *In International Conference on Machine Learning*, pages 1470–1479, 2017.
16. Matthew Hoffman, Eric Brochu, and Nando de Freitas. Portfolio allocation for bayesian optimization. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 327–336. AUAI Press, 2011.
17. Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization*, pages 507–523. Springer, 2011.
18. Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.
19. Donald R Jones, Cary D Perttunen, and Bruce E Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.
20. R Kampmann and R Wagner. Kinetics of precipitation in metastable binary alloys-theory and applications to cu-1.9 at% ti and ni-14 at% al. In *Decomposition of Alloys: The Early Stages, Proceedings of the 2 nd Acta-Scripta Metallurgica Conference*, pages 91–103, 1983.
21. Trung Le, Khanh Nguyen, Vu Nguyen, Tu Dinh Nguyen, and Dinh Phung. Gogp: Fast online regression with gaussian processes. In *IEEE 17th International Conference on Data Mining (ICDM)*, 2017.
22. Ph Lequeu, KP Smith, and A Daniélou. Aluminum-copper-lithium alloy 2050 developed for medium to thick plate. *Journal of Materials Engineering and Performance*, 19(6):841–847, 2010.
23. Cheng Li, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, and Alistair Shilton. High dimensional bayesian optimization using dropout. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2096–2102, 2017.
24. Cheng Li, Santu Rana, Sunil Gupta, Vu Nguyen, and Svetha Venkatesh. Bayesian optimization with monotonicity information. In *Workshop on Bayesian Optimization at Neural Information Processing Systems (NIPSW)*, 2017.
25. Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

26. Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. A Bayesian nonparametric approach for multi-label classification. In *Proceedings of The 8th Asian Conference on Machine Learning (ACML)*, pages 254–269, 2016.
27. Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Think globally, act locally: a local strategy for Bayesian optimization. In *Workshop on Bayesian Optimization at Neural Information Processing Systems (NIPSW)*, 2016.
28. Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Bayesian optimization in weakly specified search space. In *IEEE 17th International Conference on Data Mining (ICDM)*, 2017.
29. Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Predictive variance reduction search. In *Workshop on Bayesian Optimization at Neural Information Processing Systems (NIPSW)*, 2017.
30. Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Regret for expected improvement over the best-observed value and stopping condition. In *Proceedings of The 9th Asian Conference on Machine Learning (ACML)*, pages 279–294, 2017.
31. Vu Nguyen, Santu Rana, Sunil K Gupta, Cheng Li, and Svetha Venkatesh. Budgeted batch Bayesian optimization. In *16th International Conference on Data Mining (ICDM)*, pages 1107–1112, 2016.
32. Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.
33. Santu Rana, Cheng Li, Sunil Gupta, Vu Nguyen, and Svetha Venkatesh. High dimensional Bayesian optimization with elastic gaussian process. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2883–2891, 2017.
34. Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.
35. H Ratschek and RL Voller. Global optimization over unbounded domains. *SIAM journal on control and optimization*, 28(3):528–539, 1990.
36. Bobak Shahriari, Alexandre Bouchard-Cote, and Nando de Freitas. Unbounded Bayesian optimization via regularization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1168–1176, 2016.
37. Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
38. Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.
39. Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2171–2180, 2015.
40. Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. In *Advances in Neural Information Processing Systems*, pages 4134–4142, 2016.
41. Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022, 2010.
42. Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855. ACM, 2013.
43. Zi Wang, Bolei Zhou, and Stefanie Jegelka. Optimization as estimation with gaussian processes in bandit settings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1022–1031, 2016.
44. Ziyu Wang and Nando de Freitas. Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. *arXiv preprint arXiv:1406.7758*, 2014.
45. Dezhen Xue, Prasanna V Balachandran, John Hogden, James Theiler, Deqing Xue, and Turab Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nature communications*, 7, 2016.

Vu Nguyen is currently an Associate Research Fellow at Center for Pattern Recognition and Data Analytics (PRaDA) at Deakin University, Australia. He received his Ph.D. degrees from Deakin University, in 2015 under the supervision of Prof. Dinh Phung and Prof. Svetha Venkatesh. His research interests include Bayesian Nonparametrics, Bayesian Optimization. His international recognition includes Travel Grant Machine Learning Summer School 2011, selected participants for Heidelberg Laurate forum 2015, Finalist Awards - Track 1 and Track 5 ICPR 2016, Best Paper Runner up Award and Best Poster Award - ACML 2016, Vice Chancellor Award for Outstanding Contribution - Deakin University 2017. He gains expertise on Machine Learning and Bayesian Optimization with 20 papers published in premier machine learning venues.

Sunil Gupta is a researcher in the field of machine learning and data mining. His main interest lies in developing data-driven models for real world processes and phenomena covering both big-data and small-data problems. His recent research in optimisation using small data (Bayesian optimisation) has found applications in efficient experimental design of products and processes in advanced manufacturing such as alloy design with certain target properties, design of short nanofibers with appropriate length and thickness, and optimal setting of parameters in 3d-printers. He has published over 80 research papers including 2 Book chapters, 19 refereed journal articles, 50 fully refereed conference proceedings and 9 workshop papers with over 540 citations and an H-index of 12. His research has won several best paper awards in the field of data mining and machine learning. He is a co-inventor of 1 patent related to experimental design. He regularly serves at technical program committees of the prestigious computer science conferences.

Santu Rana is a researcher in the field of machine learning and computer vision. His broad research interests lie in devising practical machine learning algorithms for various tasks such as object recognition, mathematical optimisation and healthcare data modelling. His research in high dimensional Bayesian optimisation has been applied to efficiently design alloys with large number of elements. He has been actively conducting research in Bayesian experimental design with applications in advanced manufacturing. He has published over 52 research papers including 12 refereed journal articles, 33 fully refereed conference proceedings and 7 workshop papers with over 212 citations and an H-index of 9. He is a co-inventor of 2 patents.

Cheng Li is currently a postdoctoral fellow at Center for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Australia. He received his Ph.D. degree in 2015 at Deakin University and master degree in 2010 at Huazhong university of Science and Technology, China. His research interests lie in Bayesian optimization, graphical models and machine learning. He has published papers in top-tier conferences and journals such as ICML, IJCAI, ICDM and Scientific Report etc. He has received the ICPR2016 Best paper Finalist Award and ACML2016 Best paper Runner up Award. He is a member of 2017 Deakin Vice-Chancellor's Award for Outstanding Contribution.

Svetha Venkatesh is an ARC Australian Laureate Fellow, Alfred Deakin Professor and Director of Centre for Pattern Recognition and Data Analytics (PRaDA) at Deakin University. She was elected a Fellow of the International Association of Pattern Recognition in 2004 for contributions to formulation and extraction of semantics in multimedia data, and a Fellow of the Australian Academy of Technological Sciences and Engineering in 2006. In 2017, Professor Venkatesh was appointed an Australian Laureate Fellow, the highest individual award the Australian Research Council can bestow. Professor Venkatesh and her team have tackled a wide range of problems of societal significance, including the critical areas of autism, security and aged care. Her current research interest lies in Bayesian experimental design with applications to advanced manufacturing and healthcare. She has published over 563 research papers including 2 Books, 3 Book chapters, 156 refereed journal articles, and 402 fully refereed conference proceedings with over 15,125 citations and an H-index of 55. Additionally she is co-inventor of 3 Patents and has spun 2 start-up companies.