Impact of Invariant Objective for Order Preserving Transformation in Bayesian Optimization

Shinichi Shirakawa
Faculty of Environment and Information Science
Yokohama National University
Yokohama, Kanagawa, JAPAN
Email: shirakawa-shinichi-bg@ynu.ac.jp

***Abstract*—Bayesian optimization is a black-box optimization method, which maintains a surrogate model learned by using previously evaluated solutions and selects the next solution to be evaluated using the model. Since the Bayesian optimization method models the objective function as it is, it is not invariant under the order preserving transformation of the objective function. On the other hand, many evolutionary algorithms have that property only by using the ranking information of solutions. In this paper, we introduce two types of invariant objective function: the ranking-based objective and the Lebesgue measure-based objective, into the Bayesian optimization in order to realize the invariance property. The impact of the invariant objective function for the search performance is verified through the numerical experiment. The experimental result shows that the introduced objectives achieve the invariance for the order preserving transformation without the considerable performance deterioration in the Bayesian optimization.**

I. Introduction

Bayesian optimization [1], [2] is a global optimization method of black-box and noisy objective functions and is known as an efficient optimization algorithm particularly in an expensive scenario in which the cost of the function evaluation is so high. A surrogate model that estimates the black-box objective function is constructed and used for selecting a promising candidate solution to be evaluated. After the evaluation of the selected solution, the surrogate model is updated by adding the evaluated solution. A Gaussian process is usually adopted as the surrogate model, which can easily handle the uncertainty and noise of the objective function. Recently, the Bayesian optimization has shown effectiveness in the hyper-parameter tuning of machine learning algorithms [3]–[5], e.g., hyper-parameter tuning of the deep neural network.

In this paper, we focus on the order preserving transformation of objective function. Let us consider the two maximization problems, $\mathrm{maximize}_{x \in \mathbb{X}} f(x)$ and the transformed problem by a monotonically increasing function g, $\mathrm{maximize}_{x \in \mathbb{X}} g(f(x))$, where $\mathbb{X}$ indicates the domain of the solutions. The transformation by the function g preserves the objective rankings of solutions on the original objective function $f(x)$, i.e., $g(f(x)) < g(f(y))$ for any $x, y \in \mathbb{X}$ s.t. $f(x) < f(y)$. This means that the optimum point and the rankings of solutions do not change under such transformation. From the viewpoint of black-box optimization, the performance of optimization algorithms should not be affected by such objective transformation[1]. In some situations, the objective value is measured by logarithmic scale, $\log(f(x))$, instead of the original value $f(x)$. If the optimization algorithm is not invariant under the order preserving transformation, the performance of the optimization algorithm, such as the probability of finding the optimum and the convergence speed, must change on $\log(f(x))$ and $f(x)$. Meanwhile, the experimental results of a certain optimization algorithm in a certain optimization problem can carry over to the other order preserving transformed problems if the optimization algorithm is invariant under the order preserving transformation of objective function. In such case, we can generalize the experimental performance of the optimization algorithm not only for the specific problem but also for the set of problems. However, the Gaussian process model usually used in Bayesian optimization is not invariant under the order preserving transformation as it attempts to estimate the objective function as it is.

The invariant property for such transformation is often discussed in the evolutionary computation community. The simple way to realize the invariance property is not using the objective value as it is and only using the ranking (order) information of solutions in the optimization algorithm. Such ranking or comparison-based black-box optimization algorithms are invariant under the order preserving transformation of the objective function. Many ranking- or comparison-based methods are proposed in the context of evolutionary algorithm, such as the genetic algorithm (GA) with tournament selection, the covariance matrix evolution strategies (CMA-ES) [6], [7], and particle swarm optimization (PSO).

The ranking of a solution $\mathrm{rk}(x)$ directly links to the estimate of the quantile under the probability distribution $p(x)$ of the solutions $\{x_i\}$ [8], [9]. Let us consider the case of maximization problem, then the quantile of $f(x)$ that represents the probability of sampling a better point than x from $p(x)$ is defined by

$$q(x) = \int \mathbb{I}\{f(x) \leq f(y)\} p(y) \mathrm{d}y, \tag{1}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, and the small quantile value corresponds to the better solution. We then get the

[1] Such invariance property may not be required when we want to know the sensitivity of the objective values for the variables or estimate the concrete landscape of the objective function.

estimate of the quantile for the solution x_i by the Monte-Carlo approximation using n samples drawn from $p(x)$ as follows:

$$\hat{q}(x_i) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{f(x_i) \leq f(x_j)\} = \frac{\mathrm{rk}(x_i)}{n}, \tag{2}$$

where $\mathrm{rk}(x_i) = \#\{j, f(x_i) \leq f(x_j)\}$. Given a non-increasing function w, the original problem, maximizing $f(x)$, is reduced to the alternative problem, maximizing $w(q(x))$.

The preference weights of solutions used in the CMA-ES can be regarded as this type of objective transformation. Generally, we can interpret that the ranking-based methods measure the quality of the solutions based on the quantile $q(x)$ instead of the original objective $f(x)$. As the quantile $q(x)$ obviously depends on the probability distribution $p(x)$ of samples, the quality measure of solutions changes according to the sample distribution. Akimoto [9] has proposed a more general quality measure of solutions using Lebesgue measure, which is also invariant under the order preserving transformation of the objective function.

The goal of this paper is to introduce the invariance property for the order preserving transformation of objective function in the Bayesian optimization. Specifically, we propose to incorporate the invariant objectives, the ranking-based objective, and the Lebesgue measure-based objective [9] into the Bayesian optimization and verify their impact on the search performance through the numerical experiments.

The next section of this paper presents an overview of the Bayesian optimization. We then explain about the invariance objective functions for the order preserving transformation in Section III. In Section IV, we describe the Bayesian optimization algorithm with invariant objective functions. Next, in Section V, we verify the impact of the invariance objective functions in the Bayesian optimization through the numerical experiment. Finally, in Section VI, we describe our conclusion and future work.

II. Bayesian Optimization

Bayesian optimization [1], [2] is one of the general frameworks of the black-box optimization, especially for expensive and noisy objective functions. In the Bayesian optimization, the relatively cheap surrogate model is maintained and updated after the candidate solution is evaluated. Basically, the following steps are repeated in the Bayesian optimization algorithms:

1) Optimize the acquisition function $\alpha(x|\mathcal{M}_n)$ on the surrogate model $\mathcal{M}_n$ and get the next promising solution, $x_{n+1} = \mathrm{argmax}_{x\in\mathbb{X}}\, \alpha(x|\mathcal{M}_n)$.
2) Evaluate the selected solution, $y_{n+1} = f(x_{n+1})$.
3) Add the new observation to the observed dataset, $\mathcal{D}_{n+1} = \{\mathcal{D}_n, (x_{n+1}, y_{n+1})\}$, and update the surrogate model to $\mathcal{M}_{n+1}$ using $\mathcal{D}_{n+1}$.

A Gaussian process [10] model is often used as the surrogate model of the objective function, in which the predictive distribution is represented as the Gaussian distribution. Given the n observed data $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and the covariance function $k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, the predictive distribution is given by $\mathcal{N}(\mu_{n+1}(x), \sigma^2_{n+1}(x))$, and

$$\mu_{n+1}(x) = \mu_0 + \mathbf{k}(x)^{\mathrm{T}}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mu_0\mathbf{1}), \tag{3}$$

$$\sigma^2_{n+1}(x) = k(x,x) - \mathbf{k}(x)^{\mathrm{T}}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}(x), \tag{4}$$

where $\mathbf{K}$ is $n \times n$ covariance matrix whose element is the covariance function $K_{ij} = k(x_i, x_j)$, $\mathbf{k}(x) = [k(x, x_1), \ldots, k(x, x_n)]^{\mathrm{T}}$ is the vector of covariance between x and the observed points, $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ is the vector of the observed objective values, and $\mathbf{I}$ and $\mathbf{1}$ indicate the identity matrix and the vector of all 1, respectively. Additionally, a constant bias μ_0 corresponds to the prior mean, and σ^2 is the variance of the observation noise.

The next promising solution to be evaluated is found by optimizing the acquisition function $\alpha(x|\mathcal{M}_t)$ on the surrogate model. The acquisition function generally controls the balance between exploitation and exploration, and several acquisition functions have been proposed., e.g., the expected improvement (EI) [1], the probability improvement (PI) [1], the upper confidence-bound criteria (UCB) [11], and the mutual information (MI) [12]. The expected improvement is the most commonly used acquisition function and experimentally shows the effectiveness, given by

$$\alpha_{\mathrm{EI}}(x|\mathcal{M}) = \sigma(x)\left(Z(x)\Phi(Z(x)) + \phi(Z)\right), \tag{5}$$

$$Z(x) = \frac{\mu(x) - f(x_{\mathrm{best}})}{\sigma(x)} \tag{6}$$

where x_{best} is the current best solution, and Φ and ϕ denote the standard normal cumulative distribution and density functions, respectively.

III. Invariant Objective Function for the Order Preserving Transformation

We describe two invariant objective functions for the order preserving transformation: the ranking-based objective and the Lebesgue measure-based objective.

A. Ranking-based Objective

As we described in Section I, the ranking-based method which links to the quantile estimation is the most reasonable way for obtaining the invariance property. By transforming the original objective function into the quantile-based function, the black-box optimization algorithm becomes invariant under the order preserving transformation.

The quantile-based objective function to be maximized is given by $W^p_f(x) = w(q(x))$, where $q(x)$ is the quantile function defined in Eq. (1), and w is a non-increasing function. If we use $w(z) = -z$, the alternative objective function is represented by $W^p_f(x) = -q(x)$.

We can get the estimate of $W^p_f(x_i)$ by the Monte-Carlo approximation using n samples $\{x_1, \ldots, x_n\}$ drawn from the probability distribution $p(x)$ as

$$\hat{W}^p_f(x_i) = -\frac{\mathrm{rk}(x_i)}{n}. \tag{7}$$

Note that this transformed objective function changes depending on the sample distribution $p(x)$.

B. Lebesgue Measure-based Objective

Akimoto [9] has introduced an invariant objective function for the order preserving transformation for continuous domain and derived a theoretically attractive natural gradient method in which the parameters of the Gaussian distribution generating the candidate solutions are updated (see [9] for detail).

The Akimoto's invariant objective function to be maximized is defined by

$$\begin{aligned} V_f(x) &= -\mu_{\mathrm{Leb}}^{2/d}[y : f(x) \leq f(y)] \\ &= -\left(\int \mathbb{I}\{f(x) \leq f(y)\} \mathrm{d}y \right)^{2/d}, \end{aligned} \tag{8}$$

where μ_{Leb} denotes the Lebesgue measure on $\mathbb{R}^d$. The analytical value of $V_f(x)$ cannot be obtained because we assume that the original objective function is black-box.

However, we can estimate $V_f(x)$ by the Monte-Carlo approximation using n samples $\{x_1, \ldots, x_n\}$ drawn from the probability distribution $p(x)$.

$$\hat{V}_f(x) = -\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}\{f(x) \leq f(x_i)\}}{p(x_i)} \right)^{2/d} \tag{9}$$

Note that this transformed objective function does not depend on the sample distribution $p(x)$, but we must know the probability density function $p(x)$ to estimate it.

IV. Invariant Objective Function in Bayesian Optimization

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a black-box objective function to be maximized, defined on a bounded subset $\mathbb{X} \subset \mathbb{R}^d$. In order to realize the invariance property for the order preserving transformation of the objective function, we consider optimizing the alternative invariant objective instead of the original objective function f. We can use either $W_f^p(x)$ or $V_f(x)$, explained in Section III, as the alternative objective function. In practice, the estimated values of the invariant objectives, $\hat{W}_f^p(x)$ and $\hat{V}_f(x)$, are used.

Based on the standard Bayesian optimization manner, we use the Gaussian process to model the alternative objective function and find the next promising solution to be evaluated by optimizing the acquisition function over the model. The predictive distribution $\mathcal{N}(\mu_{n+1}(x), \sigma_{n+1}^2(x))$ of the Gaussian process model for the invariant objective function is given by replacing $\mathbf{f}$ with $\hat{\mathbf{w}} = [\hat{W}_f^p(x_1), \ldots, \hat{W}_f^p(x_n)]^{\mathrm{T}}$ or $\hat{\mathbf{v}} = [\hat{V}_f(x_1), \ldots, \hat{V}_f(x_n)]^{\mathrm{T}}$ in Eq. (3). Note that the predictive variance $\sigma_{n+1}^2(x)$ is the same as in Eq. (4).

The ranking-based objective function $\hat{W}_f^p(x_i)$ is easily obtained by Eq. (7) using n collected samples $\mathcal{D}_n$. On the other hand, we have to know the probability density function of the collected samples, $p(x)$, in order to calculate the Lebesgue measure-based objective function $\hat{V}_f(x_i)$. As the current samples $\mathcal{D}_n$ are collected by the Bayesian optimization procedure, the probability density of those samples is unknown. Therefore, we should estimate the probability density function and use it for the estimation of $\hat{V}_f(x_i)$. Let $\hat{p}(x)$ be the estimated probability density function, then the approximated invariant objective function $\hat{V}_f(x_i)$ is expressed by replacing $p(x)$ with $\hat{p}(x)$ in (9). In this paper, we use the simple kernel density estimator with the Gaussian kernel. The estimated density function using n samples is represented as

$$\hat{p}(x) = \frac{1}{\sqrt{2\pi} nh} \sum_{i=1}^{n} \exp\left(-\frac{(x - x_i)^2}{2h} \right), \tag{10}$$

where h is the bandwidth of the kernel density estimator. In the experiment, the bandwidth is selected by the Scott's rule [13] for simplicity, i.e., $h = n^{-1/(d+4)}$. More sophisticated bandwidth selection or probability density estimator should be considered in future work.

The approximated invariant objective values, $\hat{W}_f^p(x_i)$ or $\hat{V}_f(x_i)$, depend on the collected samples $\mathcal{D}_n$ and change when new data are added. Therefore, the values $\hat{W}_f^p(x_i)$ and $\hat{V}_f(x_i)$ for $i = 1, \ldots, n$ have to be recalculated at every iteration.

V. Experiment

In this section, we evaluate the impact of the invariant objective functions in the Bayesian optimization using the simple benchmark functions.

A. Experimental Setting

We implement the Bayesian optimization algorithm using the pybo[2] (version 0.1) framework [14], and the setting of the experiment is based on the default setting of the framework.

The maximum number of the evaluations is $30d$ where d indicates the problem dimension, and the initial $3d$ points are sampled according to a Latin hypercube sampling. Throughout the experiment, we use the expected improvement (EI) as the acquisition function, which is commonly used and known as the efficient acquisition function.

1) The Kernel Function: We test two types of kernel functions for the covariance function in the Gaussian process model, the automatic relevance determination (ARD) squared exponential kernel (SE) and the ARD Matérn 5/2 kernel, which are commonly used in Bayesian optimization. The ARD squared exponential kernel is defined by

$$K_{\mathrm{SE}}(x, x') = \theta_0 \exp\left(-\frac{1}{2} r^2(x, x') \right), \tag{11}$$

$$r^2(x, x') = \sum_{i=1}^{d} (x_i - x_i')^2 / \theta_i^2, \tag{12}$$

and the ARD Matérn 5/2 kernel is given by

$$\begin{aligned} K_{\mathrm{M52}}(x, x') =& \theta_0 \left(1 + \sqrt{5 r^2(x, x')} + \frac{5}{3} r^2(x, x') \right) \\ & \cdot \exp\left(-\sqrt{5 r^2(x, x')} \right), \end{aligned} \tag{13}$$

where both kernels have the $d + 1$ hyper-parameters $[\theta_0, \ldots, \theta_d]^{\mathrm{T}}$.

[2] https://github.com/mwhoffman/pybo

2) The Hyper-parameter Setting of the Gaussian Process: The hyper-parameter setting affects the modeling ability of the objective function in the Gaussian process and the performance of the Bayesian optimization. Here, we write the hyper-parameters in the Gaussian process as θ consisting of the kernel amplitude θ_0, the length scales $\theta_1, \dots, \theta_d$, the prior mean μ_0, and the variance of the observation noise σ^2.

The hyper-parameters in the Gaussian process model are marginalized by the Markov Chain Monte Carlo (MCMC) method via slice sampling, implemented in the pybo framework. Namely, we use the integrated acquisition function, $\alpha(x|\mathcal{D}_n) = \int \alpha_{\mathrm{EI}}(x|\mathcal{D}_n\theta)p(\theta|\mathcal{D}_n)\mathrm{d}\theta$. In order to compute the integrated acquisition function by MCMC, we need to set the prior distributions for θ (hyperprior). Based on the default settings of the pybo framework, we use the uniform distribution for $\theta_1, \dots, \theta_d$, the log normal distribution for θ_0, the Gaussian distribution for μ_0, and the horseshoe distribution for σ.

3) Test Problems: In order to evaluate the impact of the proposed invariant objective, we use the Branin, Gramacy, and Hartmann 6 functions, which are multimodal functions commonly used as the test functions in the Bayesian optimization literatures. The definitions of the Branin f_{Branin} and Gramacy f_{Gramacy} functions are as follows:

$$f_{\mathrm{Branin}}(x) = -\left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 - 10(1 - \frac{1}{8\pi}\cos(x_1) + 10), \quad (14)$$

$$f_{\mathrm{Gramacy}}(x) = -\frac{\sin(10\pi x_1)}{2x_1} - (x-1)^4. \quad (15)$$

The dimensions of the Branin and Gramacy are 2 and 1, respectively, and the domains of the functions are $x_1 \in [-5, 10], x_2 \in [0, 15]$ for f_{Branin}, and $x_1 \in [0.5, 0.25]$ for f_{Gramacy}. The Hartmann 6 function f_{Hart6} is a six dimensional problem, and its definition can be referred in http://www.sfu.ca/~ssurjano/hart6.html.

Additionally, we consider the following two monotonically increasing functions to investigate the effect of the order preserving transformation of the objective function:

$$g_1(y) = \frac{1}{1 + \exp(-10y)} + 10^{-5}y, \quad (16)$$

$$\text{and} \quad g_2(y) = 0.05y + 0.15(\lfloor 5y \rfloor). \quad (17)$$

Figure 1 illustrates the shapes of these functions. The function g_1 is a sigmoid-like function and g_2 is a step-like function. We can create the order preserved objective using these functions, e.g., $g_1 \circ f_{\mathrm{Branin}}$ and $g_2 \circ f_{\mathrm{Gramacy}}$.

B. Result and Discussion

The experiment was conducted with the same random seed for all algorithms in each run, and the mean value of the best solutions is calculated over the 30 independent runs.

Figures 2 and 3 show the transitions of the best function value by the Bayesian optimization with the squared exponential kernel (K_{SE}) and the ARD Matérn 5/2 kernel (K_{M52}), respectively. The results of the original objective f, the transformed objectives $g_1 \circ f$ and $g_2 \circ f$, and the invariant objectives W_f^p and V_f are plotted. The function values of the original objective function f are plotted for the transformed objectives and the invariant objectives for comparison.

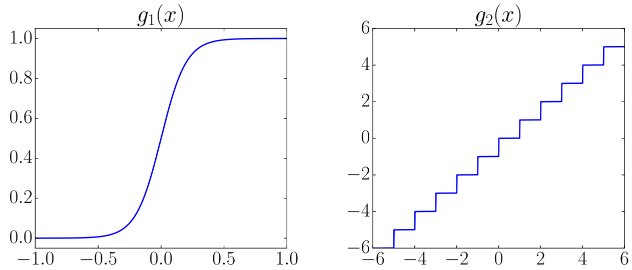

Fig. 1: The monotonically increasing functions used in the experiment, g_1 (left) and g_2 (right).

Tables I and II show the function values for the best solutions obtained on each objective by the Bayesian optimization with the kernel functions K_{SE} and K_{M52}, respectively.

We note that the results of the invariant objective functions under the order preserving transformation are exactly same because the experiment was conducted with the same random seed for all algorithm in each run. Namely, the results of W_f^p, $W_{g_1 \circ f}^p$, and $W_{g_2 \circ f}^p$ (or V_f, $V_{g_1 \circ f}$, and $V_{g_2 \circ f}$) are identical. Therefore, we only show the results on the original objective, W_f^p and V_f, for the invariant objective functions in the figures and tables.

1) Impact of the Order Preserving Transformation of the Objective Function: In both Figs. 2 and 3, the performances on the original objective function f and the transformed objective functions $g_1 \circ f$ and $g_2 \circ f$ are different, i.e., the normal Bayesian optimization method is affected by the order preserving transformation of the objective function. The performance on $g_1 \circ f_{\mathrm{Branin}}$ especially becomes worse than other transformed objective functions. On the other hand, the performances of the invariant objective functions, W_f^p and V_f, are invariant under the order preserving transformation of f. By introducing the invariant objective, we do not need to worry about the performance change by the order preserving transformation.

From the results in Tables I and II, the algorithm on the original objective function can find better solutions than that on other objectives for both the Branin and Gramacy functions. In the Hartmann 6 function, the original objective and the invariant objectives achieve almost the same quality of solution. Meanwhile, the best solutions obtained on the Lebesgue measure-based invariant objective V_f^p are competitive with those on the original objective f.

2) Comparison of the Ranking and Lebesgue Measure-based Objective: We have conducted the experiment using two invariant objective functions: the ranking-based objective and the Lebesgue measure-based objective. Interestingly, we observe that the Bayesian optimization algorithm on the

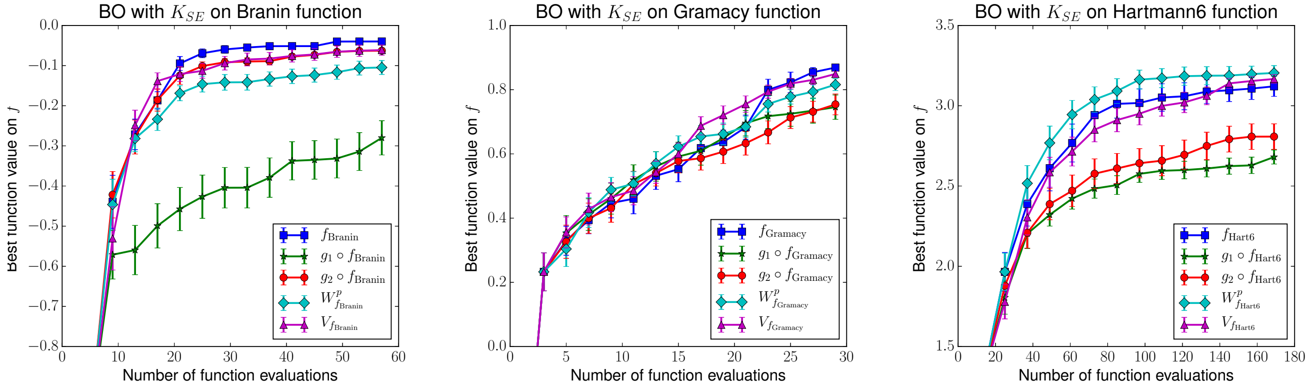

Fig. 2: Transitions of the best function value by the Bayesian optimization with the squared exponential kernel (K_{SE}) on the original objective f, the transformed objectives $g_1 \circ f$ and $g_2 \circ f$, and the invariant objectives W^p_f and V_f. The results for the Branin (left), Gramacy (center), and Hartmann 6 (right) functions are displayed. The mean values and standard errors are plotted over the 30 independent runs. For the transformed objectives and the invariant objectives, the function values of the original objective function f are plotted for comparison.

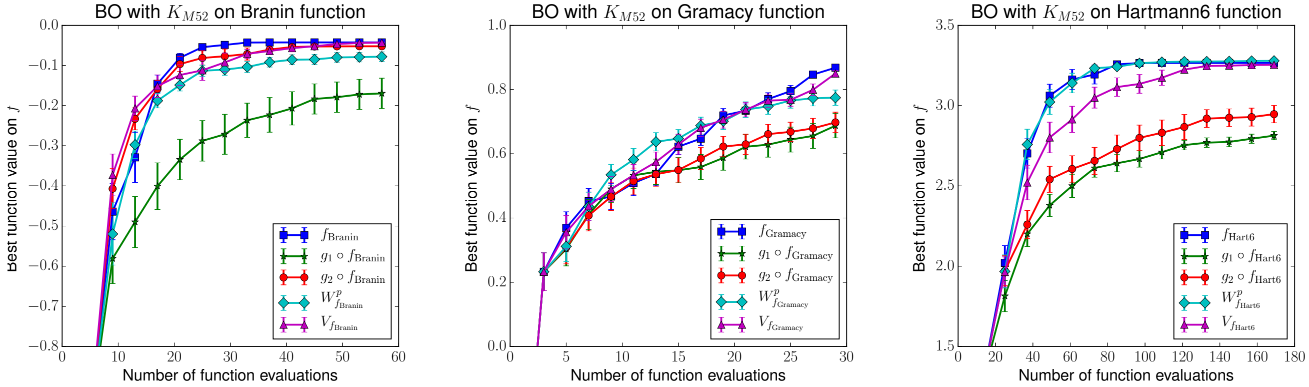

Fig. 3: Transitions of the best function value by the Bayesian optimization with the squared exponential kernel (K_{M52}) on the original objective f, the transformed objectives $g_1 \circ f$ and $g_2 \circ f$, and the invariant objectives W^p_f and V_f. The results for the Branin (left), Gramacy (center), and Hartmann 6 (right) functions are displayed. The mean values and standard errors are plotted over the 30 independent runs. For the transformed objectives and the invariant objectives, the function values of the original objective function f are plotted for comparison.

Lebesgue measure-based invariant objective can find the better solution than that on the ranking-based invariant function for both the Branin and Gramacy functions in Tables I and II. Conversely, the ranking-based objective shows the faster convergence than the Lebesgue measure-based objective in the Hartmann 6 function, but their qualities of solution are almost the same at the final iteration. The ranking-based invariant objective W^p_f depends on the distribution of the solutions, $p(x)$, which changes from iteration to iteration. This fact means that the objective function W^p_f changes at every iterations and may affect the performance decrement in the Branin and Gramacy functions.

VI. Conclusion

In this paper, we have focussed on the invariance property for the order preserving transformation of the objective function and introduced the invariant objectives into the Bayesian optimization. We have tested two types of invariant objectives, the ranking- and Lebesgue measure-based objectives, and evaluated their impact. As the analytical form of the invariant objective functions cannot be obtained, we approximate them by the Monte-Carlo. In addition, the probability density function of the current solutions is needed to compute the Lebesgue measure-based objective. Therefore, we use the simple kernel density estimation to estimate the density function. From the numerical experiment, we have confirmed that the performance of the normal Bayesian optimization changes by the order preserving transformation of the objective function, although the proposed invariant objectives do not affect that transformation. The experimental result have shown that the introduced invariant objectives achieve the invariance for the order preserving transformation without the considerable performance deterioration in the Bayesian optimization. We

TABLE I: The function values on f of the best solutions obtained on each objective function when the kernel function K_{SE} is used in the Gaussian process. The mean values and standard deviations over the 30 independent runs are reported.

Objective function	# evals.	f	$g_1 \circ f$	$g_2 \circ f$	W_f^p	V_f
Branin (f_{Branin})	60	-0.0398 ± 3.41e-6	-0.267 ± 2.32e-1	-0.0626 ± 1.87e-2	-0.105 ± 9.37e-2	-0.0490 ± 1.29e-2
Gramacy (f_{Gramacy})	30	0.869 ± 3.94e-4	0.756 ± 2.05e-1	0.759 ± 1.74e-1	0.816 ± 9.94e-2	0.849 ± 4.30e-2
Hartmann 6 (f_{Hart6})	180	3.15 ± 3.22e-1	2.68 ± 2.32e-1	2.83 ± 4.02e-1	3.22 ± 2.04e-1	3.18 ± 1.41e-1

TABLE II: The function values on f of the best solutions obtained on each objective function when the kernel function K_{M52} is used in the Gaussian process. The mean values and standard deviations over the 30 independent runs are reported.

Objective function	# evals.	f	$g_1 \circ f$	$g_2 \circ f$	W_f^p	V_f
Branin (f_{Branin})	60	-0.0424 ± 1.40e-2	-0.166 ± 2.05e-1	-0.0511 ± 1.07e-2	-0.0777 ± 6.45e-2	-0.0426 ± 6.84e-3
Gramacy (f_{Gramacy})	30	0.869 ± 1.93e-3	0.699 ± 1.99e-1	0.714 ± 1.74e-1	0.783 ± 1.18e-1	0.857 ± 3.68e-2
Hartmann 6 (f_{Hart6})	180	3.27 ± 5.91e-2	2.81 ± 1.37e-1	2.97 ± 2.92e-1	3.28 ± 5.63e-2	3.25 ± 6.47e-2

have observed that the Bayesian optimization algorithm on the Lebesgue measure-based invariant objective can find the better solution than that on the the ranking-based invariant function for the Branin and Gramacy functions.

In the Lebesgue measure-based objective, the probability density estimation is needed in order to estimate the invariant objective function. In this paper, we use only the simple kernel density estimator. Therefore, we should attempt to other probability density estimators and investigate their effect. Moreover, the impact of the invariance property under a more realistic optimization problem should be studied.

References

[1] J. Mockus, *Bayesian Approach to Global Optimization: Theory and Applications*, ser. Mathematics and its Applications. Springer, 1989.

[2] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," *arXiv preprint*, 2010. [Online]. Available: http://arxiv.org/abs/1012.2599

[3] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Proceedings of the 5th International Conference on Learning and Intelligent Optimization (LION 5)*, ser. LNCS, vol. 6683. Springer, 2011, pp. 507–523.

[4] J. Bergstra, "Algorithms for Hyper-Parameter Optimization," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 2546–2554.

[5] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2951–2959.

[6] N. Hansen, "The CMA Evolution Strategy: A Comparing Review," in *Towards a New Evolutionary Computation*, ser. Studies in Fuzziness and Soft Computing, J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 192, pp. 75–102.

[7] N. Hansen and A. Auger, "Principled Design of Continuous Stochastic Search: From Theory to Practice," in *Theory and Principled Methods for Designing Metaheustics*, ser. Natural Computing Series, Y. Borenstein and A. Moraglio, Eds. Springer Berlin Heidelberg, 2014, pp. 145–180.

[8] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen, "Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles," *arXiv preprint*, 2011. [Online]. Available: http://arxiv.org/abs/1106.3708v2

[9] Y. Akimoto, "Analysis of a natural gradient algorithm on monotonic convex-quadratic-composite functions," in *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation (GECCO 2012)*. ACM, 2012, pp. 1293–1300.

[10] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[11] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design," in *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, 2010, pp. 1015–1022.

[12] E. Contal and N. Vayatis, "Gaussian Process Optimization with Mutual Information," in *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, 2014.

[13] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 1992.

[14] M. W. Hoffman and R. Shahriari, "Modular mechanisms for Bayesian optimization," in *NIPS Workshop on Bayesian Optimization*, 2014.