A new acquisition function for Bayesian optimization based on the moment-generating function

Conference Paper · October 2017

DOI: 10.1109/SMC.2017.8122656

4 authors:

Hao Wang
Sorbonne Université
48 PUBLICATIONS **144** CITATIONS
SEE PROFILE

Bas van Stein
Leiden University
22 PUBLICATIONS **69** CITATIONS
SEE PROFILE

Michael Emmerich
Leiden University
282 PUBLICATIONS **4,097** CITATIONS
SEE PROFILE

Thomas Bäck
Leiden University
376 PUBLICATIONS **17,489** CITATIONS
SEE PROFILE

A New Acquisition Function for Bayesian Optimization Based on the Moment-Generating Function

Hao Wang
LIACS, Leiden University
Niels Bohrweg 1, Leiden
The Netherlands
Email: h.wang@liacs.leidenuniv.nl

Bas van Stein
LIACS, Leiden University
Niels Bohrweg 1, Leiden
The Netherlands
Email: b.van.stein@liacs.leidenuniv.nl

Michael Emmerich
LIACS, Leiden University
Niels Bohrweg 1, Leiden
The Netherlands
Email: m.t.m.emmerich@liacs.leidenuniv.nl

Thomas Bäck
LIACS, Leiden University
Niels Bohrweg 1, Leiden
The Netherlands
Email: t.h.w.baeck@liacs.leidenuniv.nl

***Abstract*—Bayesian Optimization or Efficient Global Optimization (EGO) is a global search strategy that is designed for expensive black-box functions. In this algorithm, a statistical model (usually the Gaussian process model) is constructed on some initial data samples. The global optimum is approached by iteratively maximizing a so-called acquisition function, that balances the exploration and exploitation effect of the search. The performance of such an algorithm is largely affected by the choice of the acquisition function. Inspired by the usage of higher moments from the Gaussian process model, it is proposed to construct a novel acquisition function based on the moment-generating function (MGF) of the improvement, which is the stochastic gain over the current best fitness value by sampling at an unknown point. This MGF-based acquisition function takes all the higher moments into account and introduces an additional real-valued parameter to control the trade-off between exploration and exploitation. The motivation, rationale and closed-form expression of the proposed function are discussed in detail. In addition, we also illustrate its advantage over other acquisition functions, especially the so-called generalized expected improvement.**

***Index Terms*—Bayesian Optimization, Acquisition function, Moment-generating function.**

I. Introduction

In many real-world optimization applications, function evaluations are very expensive. The well-known *Bayesian Optimization* algorithm is designed to optimize such problems efficiently [1]. The procedure uses a statistical model to approximate the landscape of the objective function. Such a model is usually constructed on an initial data set that is obtained by *Design of Experiments* methods (DOE) [2]. On the response surface of the model, a thorough global optimization can be performed without invoking any evaluations on the expensive objective function. However, such a naive search does not work well as the response surface sometimes differs largely from the true function landscape. This inaccuracy of the model can be quantified throughout *uncertainty quantification*. In Bayesian optimization, the *Gaussian process regression* (or Kriging) [3] is usually chosen as the model, where the theoretical variance of the prediction is given as the model uncertainty. In addition to the model estimations, the Bayesian optimization also takes the uncertainty information into account, in such a manner that an unknown solution is more preferable than the rest if the predicted value at this solution is good and / or it is associated with a relatively high uncertainty. Such a preference is translated to a real-valued function, the so-called *acquisition function* or *infill-criterion*.

The performance of such an algorithm heavily relies on the effectiveness of the acquisition function in balancing the exploration and exploitation. For instance, one acquisition can be designed to emphasize more on the model uncertainty than the prediction. Such a function is explorative and thus tends to cost more function evaluations to reach the target. However, it is also harder to get stuck in the local optimum using this function. On the contrary, it is possible to design a more exploitative acquisition function that leads to fast convergence but easily get stagnated.

Various acquisition functions have been proposed. Most of them are based on the first-order moment (expectation) from the Gaussian process. As proposed in [4], higher moments (e.g., variance, skewness) are beneficial for the acquisition function as they characterize the response surface differently and more explorative than expectation. In this paper, a novel acquisition function is developed to incorporate all the moments (to infinity), using the moment-generating function of the improvement.

This paper is organized as follows. In section II, the principle of Bayesian optimization is discussed in detail. In section III, we review the most of acquisition functions. In section IV, the MGF-based acquisition function is developed

and validated by comparing it to other functions. Finally, we conclude the paper and point out the continuation of the work.

II. Bayesian Optimization

Bayesian optimization is a *sequential* design strategy that does not require the derivatives of the objective function and is designed to solve expensive global optimization problems. Compared to alternative optimization algorithms (or other design of experiment methods), the distinctive feature of this method is the usage of a *posterior distribution* over the (partially) unknown objective function, which is obtained via *Bayesian inference*. This optimization framework is proposed by Jonas Mockus and Antanas Zilinskas, et. al. [5], [6] and later popularized by Jones et. al. [1].

In the following discussion, we assume the minimization task on some objective functions, without loss of generality (the maximization problem can be converted to minimization). Formally, the task is to approach the global minimum $\mathbf{x}^*$ of a real-valued objective function $f : S \subset \mathbb{R}^d \to \mathbb{R}$ (S is the feasible region of the search space), using a sequence of converging variables: $\lim_{n\to\infty} \mathbf{x}^{(n)} = \mathbf{x}^*,\ \mathbf{x}^{(n)} \in S \subseteq \mathbb{R}^d$, In order to achieve this, Bayesian optimization iteratively seeks for an optimal choice as the next candidate solution, obtained by optimizing a pre-defined **acquisition function** or **infill-criterion**.

A. Gaussian Process Modeling

To approximate the unknown objective function, a nonparametric regression method, Gaussian process regression [7], [8] (referred as Kriging in geostatistics [9]) is used in the Bayesian optimization. In this technique, the uncertainty of the objective function f is modeled as a probability distribution of function, which is achieved by posing a **prior** Gaussian process on it[1]. For the model training, the initial data points: $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\} \subset \mathbb{R}^d$ and the corresponding (noisy) fitness values: $\mathbf{y} = [y^{(1)}, y^{(2)}, \ldots, y^{(n)}]^\top$ are usually obtained via some sampling methods, e.g., the Latin hypercube sampling [4]. Using the data set $\mathcal{X}, \mathbf{y}$, the model updates the prior process to the so-called **posterior** process, via Bayesian inference. Specifically, in the mostly used variant, the objective function f is considered as the combination of a *centered* Gaussian Process (of zero mean) with an unknown constant trend term μ (to be estimated) [10]:

$$f \sim \mu + GP(0, k(\cdot,\cdot)),$$

where $k(\cdot,\cdot)$ is a positive definite function that computes the covariance between the function values at two distinct locations: $\mathrm{Cov}[f(\mathbf{x}), f(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}')$. Note that this GP model is also known as the *Ordinary Kriging* (OK). Moreover, in a centered Gaussian Process, any finite collection of its random variables are jointly Gaussian [3]. In this sense, the prior process on the objective function can be translated pointwise:

$$\forall \mathbf{x}, \mathbf{x}' \in S, \quad \begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}') \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} k(\mathbf{x},\mathbf{x}) & k(\mathbf{x},\mathbf{x}') \\ k(\mathbf{x},\mathbf{x}') & k(\mathbf{x}',\mathbf{x}') \end{bmatrix} \right).$$

The above expression can be extended for an arbitrarily number of points. In this paper, we choose the well-known Matérn $3/2$ kernel function for k:

$$k(\mathbf{x},\mathbf{x}') = \sigma_\varepsilon^2 \left(1 + \sqrt{3}l\right) e^{-\sqrt{3}l},\ l = \sqrt{\sum_{i=1}^{d} \theta_i \left(x_i - x_i'\right)^2},$$

where σ_ε^2 is the stationary variance of the process and θ_i's are the *hyper-parameters* of the model. Those parameter are commonly chosen through the maximum likelihood principle. Using the Bayesian inference principle to estimate the unknown trend μ, the *posterior* distribution [3] of $f(\mathbf{x})$ can be derived[2]:

$$f(\mathbf{x}) \mid \mathcal{X}, \mathbf{y}, \mathbf{x} \sim \boldsymbol{\mathcal{N}}\left(m(\mathbf{x}), s^2(\mathbf{x})\right) \tag{1}$$

$$m(\mathbf{x}) = \hat{\mu} + \mathbf{c}^\top \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \hat{\mu}\mathbf{1}_n\right) \tag{2}$$

$$s^2(\mathbf{x}) = \sigma_\varepsilon^2 - \mathbf{c}^\top \boldsymbol{\Sigma}^{-1} \mathbf{c} + \frac{\left(1 - \mathbf{c}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_n\right)^2}{\mathbf{1}_n^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_n} \tag{3}$$

$$(\mathbf{c})_i = k(\mathbf{x}, \mathbf{x}^{(i)}), \quad (\boldsymbol{\Sigma})_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), \quad \hat{\mu} = \frac{\mathbf{1}_n^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}}{\mathbf{1}_n^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_n}$$

The posterior mean function $m(\cdot)$ is the maximum a posteriori probability (MAP) estimate of the unknown f at $\mathbf{x}$ and the posterior $s(\cdot) = \sqrt{s^2(\cdot)}$ gives the standard error of the prediction $m(\cdot)$. The prediction variance / standard error quantifies the uncertainty of the model and thus plays an important role in the acquisition function.

B. Acquisition function

Given the statistical model on f, it is possible to define a "gain" $G : \mathbb{R}^d \to \mathbb{R}$ for unknown solutions. For instance, the improvement of $f(\mathbf{x})$ over the current best fitness value: $G(\mathbf{x}) \coloneqq \min(\mathbf{y}) - f(\mathbf{x})$. The acquisition function is defined as the expected gain:

$$\mathcal{A}(\mathbf{x}) = \mathbb{E}\left[G(\mathbf{x}) \mid \mathcal{X}, \mathbf{y}\right]$$

Some commonly used acquisition functions are: expected improvement, probability of improvement and upper (lower) confidence bound. The details on the acquisition function will be discussed in the next section. In each iteration, a new candidate solution $\mathbf{x}'$ is chosen by maximizing the acquisition function:

$$\mathbf{x}' = \arg\max_{\mathbf{x} \in S} \mathcal{A}(\mathbf{x})$$

After the evaluation of the new solution, we append the solution $\mathbf{x}'$ and its fitness values $f(\mathbf{x}')$ to the data set $\mathcal{X}, \mathbf{y}$. As the data set has been extended, the posterior Gaussian process is changed. Consequently, the acquisition function, which is computed from the posterior, will also evolve (locally) through the optimization.

[1] Formally, this requires to define a measurable space of the objective function.

[2] It is possible to give the posterior covariance function. See [11] for the detail.

C. The algorithm

The Bayesian optimization algorithm is summarized in Algorithm 1. The same algorithmic idea was re-advertised in the name of "Efficient Global Optimization" (EGO) by Donald R. Jones [1]. As pointed out in Jones' paper on taxonomy of global optimization methods, Bayesian optimization belongs to a broader family of optimization techniques, that is called "global optimization based on response surfaces" [12] or Sequential Model-based Optimization (MBO) [13]. In practice, some other statistical models are often exploited, e.g., Student-t processes [14], [15] and random forest [13].

Algorithm 1 Bayesian Optimization

Require: An acquisition function $\mathcal{A}$.

1 Generate the initial data set $\mathcal{X}, \mathbf{y}$ on the objective function f.
2 Construct the Gaussian process model on $\mathcal{X}, \mathbf{y}$.
3 **while** the stop criteria are not fulfilled **do**
4 Maximize the acquisition function:

$$\mathbf{x}' = \arg\max_{\mathbf{x} \in S} \mathcal{A}(\mathbf{x})$$

5 Evaluation: $y' \leftarrow f(\mathbf{x}')$.
6 Extend the data set by appending $\mathbf{x}', y'$ to $\mathcal{X}, \mathbf{y}$.
7 Re-construct the Gaussian process model of f on the augmented data set $\mathcal{X}, \mathbf{y}$
8 **end while**

III. Reviews on Acquisition functions

A lot of research effort has been put over the last decades in finding a function that provides a good balance between exploration and exploitation for various applications. In this section, we briefly review the most known and popular acquisition functions proposed in literature.

Expected Improvement (EI) It is originally proposed by Močkus [5] and utilized as the acquisition function in the standard *Efficient Global Optimization* (EGO) algorithm by Jones et. al. [1]. The EI function is highly multi-modal and tries to balance between exploration and exploitation. It is defined as follows:

$$\begin{aligned}\mathrm{EI}(\mathbf{x}) &= \mathbb{E}\left[\max(0, f_{\min} - f(\mathbf{x})) \mid \mathcal{X}, \mathbf{y}\right] \\ &= (f_{\min} - m(\mathbf{x}))\, \Phi\left(\frac{f_{\min} - m(\mathbf{x})}{s(\mathbf{x})}\right) \\ &\quad + s(\mathbf{x}) \phi\left(\frac{f_{\min} - m(\mathbf{x})}{s(\mathbf{x})}\right),\end{aligned}$$

where $f_{\min} = \min(\mathbf{y})$ is the best fitness value found so far and $m(\mathbf{x}), s(\mathbf{x})$ mean and standard deviation of the posterior Gaussian process (Eq. 1), respectively. $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function (c.d.f.) and probability density function (p.d.f.) of the standard normal random variable.

Bootstrapped Expected Improvement (B-EI) tries to correct the bias in EI as the prediction variance is known to be biased [16]. Bootstrapped EI uses parametric bootstrapping to approximate the unbiased prediction variance. Such a method is proven to be a more reliable function than EI. It, however, brings a large amount of computational time additionally.

Probability of Improvement (PI) It is an alternative function to EI [12], [17]. This function is more biased towards exploitation than exploration since it rewards the solution that are more certain (less uncertainty) to yield an improvement over the current best solution, without taking the amount of the improvement into account. The PI function is defined as:

$$\mathrm{PI}(\mathbf{x}) = P(f(\mathbf{x}) < f_{\min} \mid \mathcal{X}, \mathbf{y}) = \Phi\left(\frac{f_{min} - \hat{y}}{s}\right).$$

Lower Confidence Bound (LCB) is designed to solve the random bandit problem by balancing the exploitation and the exploration [18]. This method is adopted to the Bayesian optimization in [19] and is called Gaussian Process Upper Confidence Bound (GP-UCB). Note that its name is changed to *lower* confidence bound here as we are dealing with the minimization problem:

$$\mathrm{LCB}(\mathbf{x}; \beta) = -m(\mathbf{x}) + \sqrt{\beta} s(\mathbf{x}),$$

where β is a carefully chosen learning rate that explicitly controls the trade-off between exploitation and exploration. Obviously, a high value of β emphasizes more on the model uncertainty and thus tends to be more explorative. For the recommended setting of β, please see [19]).

Generalized Expected Improvement (GEI) [20] is a generalized form of the EI function where an additional parameter, g, is introduced to compute the g-order moment of the improvement. The larger the value of g, the more explorative the algorithm will perform and the smaller the value of g, the more locally and exploiting the function will perform.

$$\begin{aligned}\mathrm{GEI}(\mathbf{x}; g) &= \mathbb{E}\left[\max(0, f(\mathbf{x}) - f_{\min})^g \mid \mathcal{X}, \mathbf{y}\right] \\ &= s^g(\mathbf{x}) \sum_{k=0}^{g} (-1)^k \binom{g}{k} m^{g-k}(\mathbf{x}) T_k(\mathbf{x}). \quad (4)\end{aligned}$$

T_k is defined recursively for $k > 1$:

$$T_k(\mathbf{x}) = -u^{k-1}(\mathbf{x}) \phi(u(\mathbf{x})) + (k-1) T_{k-2}(\mathbf{x}), \quad (5)$$

with

$$u(\mathbf{x}) = \frac{f_{\min} - m(\mathbf{x})}{s(\mathbf{x})}, \quad T_0 = \Phi(u(\mathbf{x})), \quad T_1 = -\phi(u(\mathbf{x}))$$

The setting of the additional integer parameter g is entirely empirical. Sasena et.al. [21] proposes a "Simulated Annealing"-like approach to decrease the value of g gradually, resulting in high explorative behavior in the beginning of the optimization and more exploitative behavior after several iterations. The settings for g proposed are in the form of a look-up table where g starts at 20 and quickly goes down to 0 after iteration 35.

Multiple Generalized Expected Improvement (MGEI) and *Clustered Multiple Generalized Expected Improvement* (CMGEI) [22], are algorithms that use multiple normalized GEI functions, using different g settings, in parallel. They obtain k best local optima which are evaluated for the next iteration. The main disadvantage of this approach is the large number of evaluations required.

There are some other, less popular acquisition functions: *BayesGap* [23] and *UGap* [24], which are gap-based exploration approaches. In [25], [26], it is proposed to use portfolio strategies to select an acquisition function in each step of the Bayesian optimization.

As for optimization of the acquisition function, derivative-free Evolutionary algorithms [27] and gradient-based method (e.g., quasi-Newton's method) are often used / combined to search for the global optimum. As an alternative, Wang et. al. [28] propose to diversify the search by adapting the niching techniques to find multiple (local) optima of the acquisition function.

In [29], [30], multiple conflicting acquisition functions are considered together (e.g. PI and EI), which forms a multi-objective optimization problem naturally. Such a multi-objective treatment gives the decision makers the flexibility to choose among low-risk and / or high-gain solutions and possibly leads to parallelization of the Bayesian optimization. An alternative approach proposed by Hutter et. al. [31], instantiates multiple LCB functions by sampling several β values from an exponential distribution with the unit mean.

In the aforementioned works, Some comparisons of these acquisition functions have been performed in literature as well. For example, a conceptual comparison between various, early proposed, acquisition functions can be found in [12].

IV. MGF-based Acquisition Function

A. The proposal

In this section, we propose the new acquisition function that is constructed based on the *moment-generating function* of the improvement. Formally, the improvement is defined as [20]:

$$I(\mathbf{x}) = \begin{cases} f_{\min} - f(\mathbf{x}) & \text{if } f(\mathbf{x}) < f_{\min}, \\ 0 & \text{otherwise.} \end{cases}$$

$f_{\min}$ stands for the fitness value of the best solution found so far. When modeling the objective function as a Gaussian process: $f(\mathbf{x}) \mid \mathcal{X}, \mathbf{y} \sim \mathcal{N}(m(\mathbf{x}), s^2(\mathbf{x}))$ (Eq. 1), the distribution of $I(\mathbf{x})$ is known as *rectified Gaussian*[3], whose density function is written as:

$$p_{I(\mathbf{x})}(u) = \begin{cases} \Phi\left(\frac{m(\mathbf{x})-f_{\min}}{s(\mathbf{x})}\right)\delta(u) & u < f_{\min}, \\ \frac{1}{s(\mathbf{x})\sqrt{2\pi}} \exp\left(-\frac{(s-(f_{\min}-m(\mathbf{x})))^2}{2s^2(\mathbf{x})}\right) & \text{otherwise.} \end{cases}$$

$\delta(\cdot)$ is the well-known Dirac delta function. Most of the acquisition functions reviewed in the previous section are constructed to summarize the statistical property of the improvement over the current best solution. For instance, the Generalized Expected Improvement calculates the moments about the origin of $I(\mathbf{x})$.

Following the proposal to use the higher moments in the Bayesian Optimization, it is proposed to develop an acquisition function based on the *Moment-Generating Function* (MGF) of the improvement. Loosely speaking, given the existence of the moment-generating function, it can be expanded as a Taylor series, whose terms are proportional to all the moments (to the infinite order) of the improvement. Therefore, such a function is considered as combination of all the moments. Formally, MGF of the improvement $I(\mathbf{x})$ is an alternative way to give its probability distribution and it is defined as:

$$\forall t \in \mathbb{R}, \quad M(\mathbf{x}, t) := \mathbb{E}\left[e^{tI(\mathbf{x})}\right] = \int_{-\infty}^{\infty} e^{tu} p_{I(\mathbf{x})}(s)\,\mathrm{d}u$$

Moreover, the moment-generating function can be calculated using the density function of $I(\mathbf{x})$:

$$\begin{aligned} M(\mathbf{x}, t) &= \Phi\left(\frac{f_{\min} - m'(\mathbf{x})}{s(\mathbf{x})}\right) \exp\left((f_{\min} - m(\mathbf{x}))\, t + \frac{s^2(\mathbf{x})t^2}{2}\right) \\ &\quad + 1 - \Phi\left(\frac{f_{\min} - m(\mathbf{x})}{s(\mathbf{x})}\right) \\ m'(\mathbf{x}) &= m(\mathbf{x}) - s^2(\mathbf{x})t \end{aligned} \tag{6}$$

This function has a closed form and is well-defined for all $t \in \mathbb{R}$. From a different perspective, the Taylor expansion of this function is:

$$\begin{aligned} M(\mathbf{x}, t) &= 1 + t\mathbb{E}[I(\mathbf{x})] + \frac{t^2}{2!}\mathbb{E}[I^2(\mathbf{x})] + \frac{t^3}{3!}\mathbb{E}[I^3(\mathbf{x})] + \cdots \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}[I^n(\mathbf{x})]. \end{aligned} \tag{7}$$

Note that, for an arbitrary distribution, the above series might not converge for all the $t \in \mathbb{R}$, even if all the moments exist. When treating $t^n/n!$ as the weight for each moment $\mathbb{E}[I^n]$, this function can also be considered as a linear combination of the moments, where the weights are controlled by variable t. In addition, it is possible to normalize the weights by observing the fact that: $\sum_{n=0}^{\infty} \frac{t^n}{n!} = e^t$, which converges for all $t \in \mathbb{R}$. Thus, the normalized MFG function[4] is obtained by dividing it by e^t. The additional parameter t controls the trade-off between exploration and exploitation of the global search. To visualize this, multiple sets of weights are plotted in Fig. 1 by varying t. According to the figure, a low value of t (e.g., $t < 1$) assigns more weights to the lower moments (e.g., the expected improvement), rendering the search process mainly exploitative. As for the higher values of t, the "center" (mean) of the weight distribution is t and dispersion (variance) is also increasing with respect to t. This indicates that more higher moments of the improvement are taken into account when t increases and therefore the search tends to be explorative.

[3]Do not confuse the rectified Gaussian with the so-called truncated Gaussian distribution.

[4]In fact, the normalized weight, $\frac{t^n}{e^t n!}$ is exactly the probability mass function of the Poisson distribution.

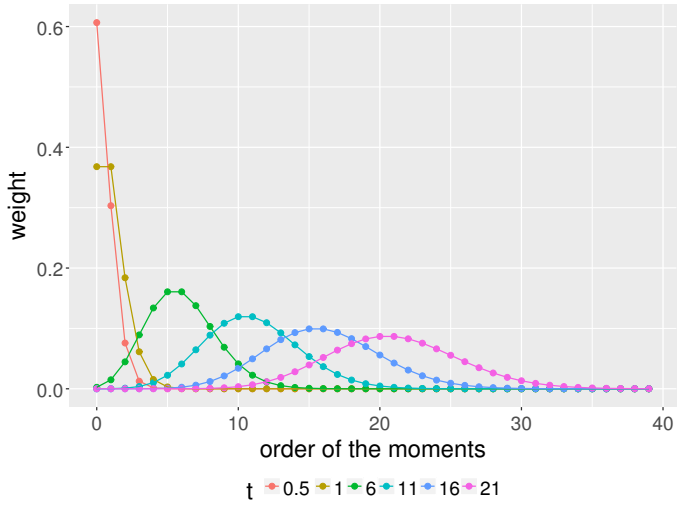


Fig. 1. Distribution of the combination weights in the normalized moment-generating function by varying the t value from 0.5 to 21.

Finally, it is proposed to incorporate the Probability of Improvement (PI) in the proposed acquisition function. This is achieved by treating PI as the "zero-order" moment of $I(\mathbf{x})$ and replacing the constant 1 in Eq. 7. Putting all the considerations together, the proposed MGF-based acquisition function is:

$$\begin{aligned}
&\mathcal{M}(\mathbf{x};t) \\
&= \mathrm{PI}(\mathbf{x}) + \frac{t}{e^t}\mathbb{E}[I(\mathbf{x})] + \frac{t^2}{2!e^t}\mathbb{E}[I^2(\mathbf{x})] + \frac{t^3}{3!e^t}\mathbb{E}[I^3(\mathbf{x})] + \cdots \\
&= \frac{M(\mathbf{x},t) - 1 + \mathrm{PI}(\mathbf{x})}{e^t} \\
&= \Phi\left(\frac{f_{\min} - m'(\mathbf{x})}{s(\mathbf{x})}\right) \exp\left(\left(f_{\min} - m(\mathbf{x}) - 1\right)t + \frac{s^2(\mathbf{x})t^2}{2}\right)
\end{aligned} \tag{8}$$

where $m'(\mathbf{x})$ is defined in Eq. 6. In order to align with existing work [30] on using gradient-based optimization techniques for the acquisition function, the gradient of the proposed acquisition function is given as:

$$\begin{aligned}
\nabla\mathcal{M}(\mathbf{x};t) = C\Bigg[&\Phi\left(\frac{f_{\min} - m'(\mathbf{x})}{s(\mathbf{x})}\right)\left(t^2 s(\mathbf{x})\nabla s(\mathbf{x}) - t\nabla m\right) \\
&- \phi\left(\frac{f_{\min} - m'(\mathbf{x})}{s(\mathbf{x})}\right)\left(\frac{\nabla m'}{s(\mathbf{x})} + \frac{f_{\min} - m'}{s^2(\mathbf{x})}\nabla s(\mathbf{x})\right)\Bigg] \\
\nabla m'(\mathbf{x}) &= \nabla m - 2ts(\mathbf{x})\nabla s(\mathbf{x}) \\
C &= \exp\left(\left(f_{\min} - m(\mathbf{x}) - 1\right)t + \frac{s^2(\mathbf{x})t^2}{2}\right)
\end{aligned}$$

The expression for the gradient $\nabla m(\mathbf{x})$ and $\nabla s(\mathbf{x})$ can be found in [30].

B. Comparison to GEI

The proposed acquisition function is designed to exploit the higher moments of the improvement. Similarly, the Generalized Expected Improvement (Eq. 4) is proposed for the same purpose. The main advantages of the newly proposed acquisition function $\mathcal{M}$ over GEI are:

1) $\mathcal{M}$ combines all the moments using a weight distribution instead of using one moment each time in GEI. This leads to a much smaller "change" in the acquisition function when tuning the addition parameter t.
2) It has a simple closed-form expression, in contrast to a recursive formula (Eq. 5) for GEI. As a result, it is obvious to see that $\mathcal{M}$ is computationally less expensive than GEI. The gradient of $\mathcal{M}$ can also be easily calculated.
3) The extra parameter t that balances the exploration and exploitation, takes continuous values while the parameter g in GEI is a integer variable. Consequently, when tuning this additional parameter, the effect can be more smooth: a more realistic cooling schedule can be applied. For example, the common exponential decay can be applied: $t \leftarrow (1-\alpha)t, \quad 0 < \alpha < 1$.

The difference between the MGF-based acquisition function and GEI is illustrated in Fig. 2. On the 1-d Ackley function, a Gaussian process model is built on 6 uniformly distributed samples (black dots) in $[-5, 7]$. 11 MGF-based acquisition functions (on the left sub-figure) are created using a log-scaled t value from roughly 0.1 to 3 while 11 GEI functions (on the right) are depicted with g from 0 to 10. The maximum of the acquisition function (the black stars) are sampled in the next iteration. Comparing the spread of those maximum points, it is obvious that: when g increases, those maxima are getting close to each other in the GEI scenario and thus become indistinguishable. This means the GEI functions are, in fact, indifferent when the parameter g increases. However, for the MGF-based functions, the maxima still show significant differences when t is large. Therefore, the effective range of the g value is much narrower than that of t.

V. Conclusion

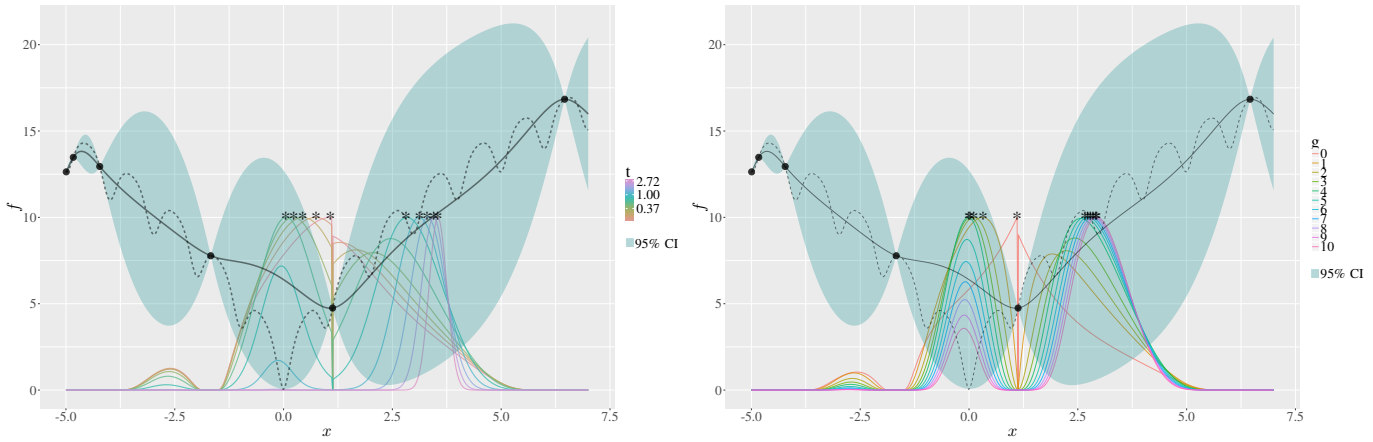
In the paper, a novel acquisition function is developed for Bayesian optimization. It is constructed from the moment-generating function of the improvement. The Taylor expansion of this function shows that it is a weighted combination of all the moments (infinitely many) of the improvement. Therefore, it refines the so-called generalized expected improvement.

Moreover, the proposed acquisition function introduces an real-valued parameter to explicitly control the exploration / exploitation trade-off. Then it is straightforward to apply a "cooling" schedule on this parameter, making the Bayesian optimization more explorative in the beginning and more exploitative in the long run.

For the future work, a systematic empirical study is required to investigate the impact of the proposed acquisition function. The optimal cooling strategy should also be considered.

References

[1] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.
[2] D. C. Montgomery, "Design and analysis of experiments," 1991.



(a) MGF-based acquisition function

(b) Generalized Expected Improvement

Fig. 2. On the 1-D Ackley function, the acquisition functions are plotted by varying parameters t and g. The objective function is shown as the dashed curve while the GP model is presented in solid curve, which is trained on the black points. The maximum of the acquisition function is indicated by the black stars. The shaded area illustrated the 95% confidence interval. All the acquisition function values are normalized and rescaled to $[0, 10]$.

[3] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, ser. Adaptive computation and machine learning series. University Press Group Limited, 2006.

[4] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979. [Online]. Available: http://www.jstor.org/stable/1268522

[5] J. Močkus, "On bayesian methods for seeking the extremum," in *Optimization Techniques IFIP Technical Conference*. Springer, 1975, pp. 400–404.

[6] J. Mockus, *Bayesian approach to global optimization: theory and applications*. Springer Science & Business Media, 2012, vol. 37.

[7] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and Analysis of Computer Experiments," *Statistical Science*, vol. 4, no. 4, pp. 409–423, 1989.

[8] T. Santner, B. Williams, and W. Notz, *The Design and Analysis of Computer Experiments*, ser. Springer Series in Statistics. Springer, 2003.

[9] D. G. Krige, "A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand," *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, vol. 52, no. 6, pp. 119–139, Dec. 1951.

[10] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[11] D. Ginsbourger, R. Le Riche, and L. Carraro, "Kriging is well-suited to parallelize optimization," in *Computational Intelligence in Expensive Optimization Problems*. Springer, 2010, pp. 131–162.

[12] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[13] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *International Conference on Learning and Intelligent Optimization*. Springer, 2011, pp. 507–523.

[14] B. J. Williams, T. J. Santner, and W. I. Notz, "Sequential design of computer experiments to minimize integrated response functions," *Statistica Sinica*, pp. 1133–1152, 2000.

[15] A. Shah, A. Wilson, and Z. Ghahramani, "Student-t processes as alternatives to gaussian processes," in *Artificial Intelligence and Statistics*, 2014, pp. 877–885.

[16] J. P. C. Kleijnen, W. van Beers, and I. van Nieuwenhuyse, "Expected improvement in efficient global optimization through bootstrapped kriging," *Journal of Global Optimization*, vol. 54, no. 1, pp. 59–73, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10898-011-9741-y

[17] A. Žilinskas, "A review of statistical models for global optimization," *Journal of Global Optimization*, vol. 2, no. 2, pp. 145–153, 1992.

[18] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[19] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design," *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 1015–1022, 2010. [Online]. Available: http://arxiv.org/abs/0912.3995

[20] M. Schonlau, W. J. Welch, and D. R. Jones, "Global versus local search in constrained optimization of computer models," *Lecture Notes-Monograph Series*, pp. 11–25, 1998.

[21] M. J. Sasena, P. Papalambros, and P. Goovaerts, "Exploration of metamodeling sampling criteria for constrained global optimization," *Engineering optimization*, vol. 34, no. 3, pp. 263–278, 2002.

[22] W. Ponweiser, T. Wagner, and M. Vincze, "Clustered Multiple Generalized Expected Improvement: A novel infill sampling criterion for surrogate models," *2008 IEEE Congress on Evolutionary Computation, CEC 2008*, no. April, pp. 3515–3522, 2008.

[23] M. W. Hoffman, B. Shahriari, and N. de Freitas, "On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning," *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, vol. 33, pp. 365–374, 2014.

[24] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *Advances in Neural Information Processing Systems*, 2012, pp. 3212–3220.

[25] M. Hoffman, E. Brochu, and N. D. Freitas, "Portfolio Allocation for Bayesian Optimization," *Conference on Uncertainty in Artificial Intelligence*, pp. 327–336, 2011.

[26] R. K. Ursem, "From expected improvement to investment portfolio improvement: Spreading the risk in kriging-based optimization," in *International Conference on Parallel Problem Solving from Nature*. Springer, 2014, pp. 362–372.

[27] T. Bäck and H.-P. Schwefel, "An overview of evolutionary algorithms for parameter optimization," *Evol. Comput.*, vol. 1, no. 1, pp. 1–23, Mar. 1993. [Online]. Available: http://dx.doi.org/10.1162/evco.1993.1.1.1

[28] H. Wang, T. Bäck, and M. Emmerich, "Multi-point efficient global optimization using niching evolution strategy."

[29] B. Bischl, S. Wessing, N. Bauer, K. Friedrichs, and C. Weihs, "Moi-mbo: Multiobjective infill for parallel model-based optimization," in *International Conference on Learning and Intelligent Optimization*. Springer, 2014, pp. 173–186.

[30] H. Wang, M. Emmerich, and T. Back, "Balancing risk and expected gain in kriging-based global optimization," in *Evolutionary Computation (CEC), 2016 IEEE Congress on*. IEEE, 2016, pp. 719–727.

[31] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Parallel algorithm configuration." *LION*, vol. 6, pp. 55–70, 2012.