Advancing Bayesian Optimization: The Mixed-Global-Local (MGL) Kernel and Length-Scale Cool Down

Kim P. Wabersich
Machine Learning and Robotics Lab
University of Stuttgart, Germany
wabersich@kimpeter.de

Marc Toussaint
Machine Learning and Robotics Lab
University of Stuttgart, Germany
marc.toussaint@informatik.uni-stuttgart.de

Abstract

Bayesian Optimization (BO) has become a core method for solving expensive black-box optimization problems. While much research focussed on the choice of the acquisition function, we focus on online length-scale adaption and the choice of kernel function. Instead of choosing hyperparameters in view of maximum likelihood on past data, we propose to use the acquisition function to decide on hyperparameter adaptation more robustly and in view of the future optimization progress. Further, we propose a particular kernel function that includes non-stationarity and local anisotropy and thereby implicitly integrates the efficiency of local convex optimization with global Bayesian optimization. Comparisons to state-of-the art BO methods underline the efficiency of these mechanisms on global optimization benchmarks.

1 Introduction

Bayesian Optimzation (BO) became an almost ubiquitous tool for general black-box optimization with high function evaluation cost. A BO algorithm is in principle characterized by two choices: 1) What is the prior over the objective function? 2) Given a posterior, what is the decision theoretic criterion, the so-called acquisition function, to choose the next query point? Previous research has extensively focussed on the second question. In this paper we rather focus on the first question, the choice of model or prior over the objective function. Clearly, from the purely Bayesian stance the prior must be given and is not subject to discussion. However, there are a number of reasons to reconsider this:

Choice of Hyperparameters: In practice, choosing the hyperprior online (e.g. using leave-one-out cross-validation (LOO-CV) on the so-far seen data) is prone to local optima and may lead to significant inefficiency w.r.t. the optimization process as already mentioned in [1]. In this paper we take the stance that if one chooses a point estimate for the hyperprior online, then maximum likelihood only on the seen data is *not* an appropriate model selection criterion. Instead, we should choose the hyperprior so as to accelerate the optimization process.

Choice of kernel function: The squared-exponential kernel is the standard choice of prior. However, this is in fact a rather strong prior as many relevant functions are heteroscedastic (have different length-scales in different regions) and have various local optima, each with different non-isotropic conditioning of the Hessian at the local optimum. Only very few preliminary experiments on heteroscedastic and non-isotropic models have been reported [2, 3]. In this paper we propose a novel type of kernel function with the following in mind. Classical model-based optimization of convex black-box functions [4, Section 8] is extremely efficient *iff* we know the function to be convex. Therefore, for the purpose of optimization we may presume that the objective function

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

has local convex polynomial regions, that is, regions in which the objective function is convex and can reasonably be approximated with a (non-isotropic) 2nd-order polynomial, such that within these regions, quasi-Newton type methods converge very efficiently. To this effect we propose the Mixed-Global-Local (MGL) kernel, which expresses the prior assumption about local convex polynomial regions, as well as automatically implying a local search strategy that is analogous to local model-based optimization. Effectively, this choice of kernel integrates the efficiency of local model-based optimization within the Bayesian optimization framework.

1.1 Background

Algorithm 1 General Bayesian optimization

```
 1: procedure GBO(objective $f$, $\mathcal{GP}(c_\mu, k)$ , max. Iterations $N$, acquisition function $\alpha$)
 2:     init $X_0 = \{\boldsymbol{x}_{01}, ..., \boldsymbol{x}_{0N_i}\}$, $\boldsymbol{x}_0 \in \mathcal{D}$
 3:     init $\boldsymbol{y}_0 = [f(\boldsymbol{x}_{01}), ..., f(\boldsymbol{x}_{0N_i})]^T$
 4:     $n \leftarrow 1$
 5:     for $n \leq N$ do
 6:         perform model adaption
 7:         $\boldsymbol{x}_n = \text{argmin}_{\boldsymbol{x} \in \mathcal{D}} \alpha_n(\boldsymbol{x})$
 8:         $X_n \leftarrow \{\boldsymbol{x}_n\} \cup X_{n-1}$
 9:         $\boldsymbol{y}_n \leftarrow \{f(\boldsymbol{x}_n)\} \cup \boldsymbol{y}_{n-1}$
10:         $n = n + 1$
11:     end for
12:     $n^* \leftarrow \text{argmin}_n \, y_n \in \boldsymbol{y}_N$
13:     return $\boldsymbol{x}_{n^*}$                ▷ best observation
14: end procedure
```

We consider the black-box optimization problem

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x}) \tag{1}$$

with an objective $f : \mathcal{D} \to \mathbb{R}$ that maps a hypercube

$$\mathcal{D} = \{\boldsymbol{x} \in \mathbb{R}^d \mid x_i \in [0,\ 1] \subset \mathbb{R}, i = 1, 2, .., d\} \tag{2}$$

to real numbers. Therefor we use a Gaussian process (GP) [5] prior over f with constant prior mean function $\mu = c_\mu$. Together with a covariance (kernel) function $k(\boldsymbol{x}, \boldsymbol{x}')$ we write $\mathcal{GP}(c_\mu, k)$ for the prior GP in short. A very common choice of kernel is the squared exponential (SE) kernel

$$k_{\text{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-0.5 \frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{l^2}\right). \tag{3}$$

The GP model assumption about f builds the basis for many BO algorithms. A general prototype for such an algorithm is given in Alg. 1, where α_n is the algorithm specific acquisition function. For experiments we will use the well known and theoretical extensively studied Expected Improvement (EI) [6] acquisition function. In this work we particurlarly address the choice of k and the model adaption in Alg. 1, line 6.

1.2 Related work

In [3] they introduce the idea of *local* length-scale adaption based on maximizing the acquisition function (EI) value, which is not efficient as they say (and different to the cool down we propose). Nevertheless we endorse the underlying idea, since it is related to our motivation.

On the model side there are several ideas which yield non-isotropic models by building an ensemble of local isotropic kernels, e.g. based on trees [7]. We however introduce a specific kernel rather than a concept of combining kernels or Gaussian processes taylored for improving BO. Another approach was presented in [8] which relies on a non-stationary input transformation combined with a stationary kernel for improving performance in case of non-stationary objective functions.

There are also concepts regarding locally defined kernels, e.g. [9]. The idea of [2] is somehow closely related to ours, because they use a local and a global kernel function, which is a great approach, as we believe. They parametrize the location of the local kernel as well as the respective parameters. Consequently they end up with a large number of hyperparameters which makes model selection very difficult. In constrast to their work we are able to gain comparable or better performance in well-known benchmarks. At the same time we overcome the problem of many hyperparameters by a separated, efficient algorithm for determining the location of local minimum regions. Furthermore we use a non-isotropic kernel for better fitting local minimum regions.

2 Alpha-ratio cool down

In this section we address length-scale adjustment of an isotropic kernel during the optimization process as part of the general BO Algorithm (Alg. 1, Line 6).

Let l_{n-1} be the length-scale used in the previous iteration. In our approach we want to decide whether to reuse the same length-scale or decrease it to a specific smaller length-scale $\tilde{l}_n < l_{n-1}$ in iteration n. In our experiments we will choose $\tilde{l}_n = \max(l_{n-1}/2, \bar{l}_n)$, where

$$\bar{l}_n(d, \bar{c}) = \sqrt{-\frac{1}{2\log(\bar{c})}} \left(\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{3}{2})} \pi^{0.5(1-d)} \frac{1}{n} \right)^{\frac{1}{d}} \tag{4}$$

is a hard lower bound that encodes a minimal correlation $\bar{c}$ between sampling points in case of an approximate uniform sphere packed data set X_n. Since Alg. 1 will not select data in this "explorative sense" the minimal correlation $\bar{c}$ will be violated and thus serves as lower bound. We propose to use the aquisition function as a criterion for the decision to decrease the length-scale. Let

$$\alpha_{r,n} := \frac{\alpha^*(\tilde{l}_n)}{\alpha^*(l_{n-1})} \tag{5}$$

be the alpha-ratio, where $\alpha^*(l) = \min_{\boldsymbol{x} \in \mathcal{D}} \alpha_n(\boldsymbol{x}; l)$ is the optimal aquisition value when using length-scale l. In typical situations we expect that $\alpha_{r,n} > 1$ because the reduced length-scale $\tilde{l}_n$ leads to larger posterior variance, which typically leads to larger aquisition values, i.e., more chances for progress in the optimization process. We turn this argument around: if $\alpha_{r,n}$ is not *substantially* larger than 1, then choosing the smaller length-scale $\tilde{l}_n$ does not yield substantially more chances for progress in the optimization process. In this case, as a smaller length-scale has higher risk of overfitting, we decide to stick to the old length-scale l_{n-1}.

In summary, in our *alpha-ratio (AR) cool down* for length-scale adaption we have a fixed threshold $\bar{\alpha}_r > 1$ and choose $l_n = \tilde{l}_n$ as new length-scale if $\alpha_{r,n} > \bar{\alpha}_r$, and $l_n = l_{n-1}$ otherwise.

3 Mixed-global-local kernel

We assume that each local (global) optimum $\boldsymbol{x}_i^*$ of (1) is within a neighbourhood $\mathcal{U}_i(\boldsymbol{x}_i^*)$ that can be approximated by a positive definite quadratic function. More precisely:

Definition 1. *Given a data set $D = \{(\boldsymbol{x}_i, y_i)\}$, we call a convex subset $\mathcal{U} \subset \mathcal{D}$ a convex neighborhood if the solution of the regression problem*

$$\{\beta_0^*, \boldsymbol{\beta}_1^*, B^*\} = \underset{\beta_0, \boldsymbol{\beta}_1, B}{\operatorname{argmin}} \sum_{k: \boldsymbol{x}_k \in \mathcal{U}} \left[(\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_k + \frac{1}{2} \boldsymbol{x}_k^T B \boldsymbol{x}_k) - y_k \right]^2 , \tag{6}$$

($\boldsymbol{x}_k \in \mathcal{U}$ the data points in $\mathcal{U}$) has a positive definite Hessian B.

If we are given a set $\{\mathcal{U}_i\}$ of convex neighborhoods that are pair-wise disjoint we define the following kernel function:

Definition 2. *The Mixed-Global-Local (MGL) kernel is given by*

$$k_{MGL}(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} k_q(\boldsymbol{x}, \boldsymbol{x}'), \ \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{U}_i, \\ k_s(\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{x} \notin \mathcal{U}_i, \boldsymbol{x}' \notin \mathcal{U}_j \\ 0, \ else \end{cases} \tag{7}$$

for any i, j, where k_s is a stationary-isotropic kernel [5] and

$$k_q(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + 1)^2 \tag{8}$$

the quadratic kernel.

This kernel is heteroscedastic in the sense that the quadratic kernels in the convex neighborhood implies fully different variances than the “global” stationary-isotropic kernel around the neighborhoods.

For determining $\mathcal{U}_i$ we discretize the search space using the samples as centers for k-nearest-neighbor (kNN) search. As soon as a kNN tuple of samples satisfy Def. 1, we get a ball shaped local minimum region candidate. We add a local convergence criteria, that is, the minimum of a local region must have a minimum distance $\epsilon > 0$ to any sample. At the end we remove all region candidates that overlap with better regions.

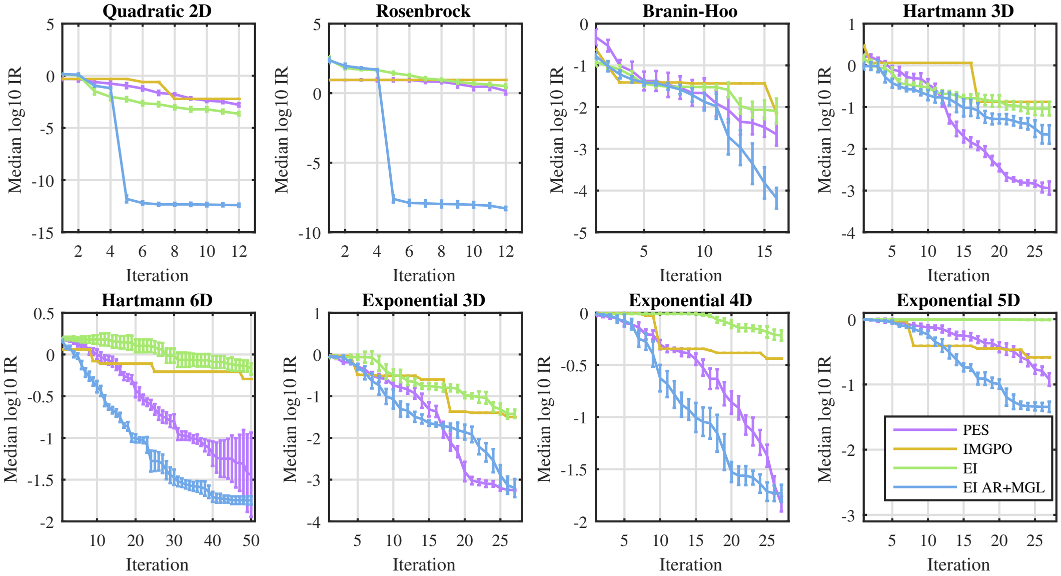

Figure 1: Comparison of recent Bayesian optimization algorithms with synthetic test functions.

4 Empirical results

For all tests we choose the following configurations: We set $\bar{c} = 0.2$, $\bar{\alpha}_r = 1.5$. For the MGL-kernel (7) we take the SE kernel (3) for k_s. We estimated the observation variance σ_f^2 in (3) and the constant mean of the prior GP via maximum likelihood and scaled the observation variance down by factor 100 for consistency with the quadratic part of (7) if any local region is detected. For computing Alg. 1 line 7 we first solved the minimization using the k_s kernel of (7) and compared it with the results of the minimization problems using the k_q kernel for each local minimum region $\mathcal{U}_i$ since all the regions for the different kernel parts are disjoint. We used three samples as initial design set, chosen by latin hypercube sampling.

In the following we will often refer to an *optimal* choice of hyperparameters. By this we mean that 1000 random samples from the respective objective function are taken. On this data an exhaustive LOO-CV is used to select the length-scale, and max-likelihood to select the prior variance σ_f^2 and mean-prior c_μ.

In Fig. 1 we report on results using several synthetic benchmark functions. Shown are predictive entropy search (PES) [10] (which treats hyperparameters in a Bayesian way in the acquisition function), infinite metric GP optimization (IMGPO) (which uses a Bayesian update for hyperparameters in each iteration), classical EI with optimal hyperparameters, and EI using our alpha-ratio model adaption and the MGL-kernel (EI AR + MGL). For all performance tests where we show the log10 median performance (Immidiate Regret (IR)), we made 32 runs and estimated the median variance via bootstrapping. The errorbars indicate one times the standard deviation. In addition to commonly considered benchmark functions (Rosenbrock, Branin-Hoo, Hartmann3D, Hartmann 6D) taken from [11], we show a simple quadratic function in the interval $[-2, 2]^2$ and an exponential function of the form $f_{\text{exp}}(\boldsymbol{x}) = 1 - \exp(\boldsymbol{x}^T C \boldsymbol{x})$ with $C := \text{diag}([10^{0/(d-1)}, 10^{1/(d-1)}, .., 10^{(d-1)/(d-1)}])$ on the same interval in respective dimensions d.

The MGL-kernel outperforms significantly in case of the quadratic and the more quadratic like Rosenbrock objective. Also for Branin-Hoo, Hartmann 6D and Exponential 5D our method significantly outperforms existing state-of-the-art Bayesian optimization methods. In case of Hartmann 3D, PES turns out to work better. Nevertheless we want to emphasize the outstanding improvement compared to plain EI with optimal hyperparameters in every test case.

5 References

[1] Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

[2] Ruben Martinez-Cantin. Locally-biased Bayesian optimization using nonstationary Gaussian processes. In *Neural Information Processing Systems (NIPS) workshop on Bayesian Optimization*, 2015.

[3] Hossein Mohammadi, Rodolphe Le Riche, and Eric Touboul. Small ensembles of kriging models for optimization. *arXiv preprint arXiv:1603.02638*, 2016.

[4] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[5] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[6] Adam D Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(Oct):2879–2904, 2011.

[7] John-Alexander M Assael, Ziyu Wang, Bobak Shahriari, and Nando de Freitas. Heteroscedastic treed bayesian optimisation. *arXiv preprint arXiv:1410.7172*, 2014.

[8] Jasper Snoek, Kevin Swersky, Richard S Zemel, and Ryan P Adams. Input warping for bayesian optimization of non-stationary functions. In *ICML*, pages 1674–1682, 2014.

[9] Andreas Krause and Carlos Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *Proceedings of the 24th international conference on Machine learning*, pages 449–456. ACM, 2007.

[10] 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.

[11] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved September 21, 2016, from `http://www.sfu.ca/~ssurjano`, 2016.