Correcting boundary over-exploration deficiencies in Bayesian optimization with virtual derivative sign observations

Eero Siivola[1], **Aki Vehtari**[1], **Jarno Vanhatalo**[2], **Javier González**[3,*]
[1]Aalto University, Dept. of Computer Science, {eero.siivola, aki.vehtari}@aalto.fi
[2]Univ. of Helsinki, Dept. of Math. and Stat.,
and Dept. of Biosciences, jarno.vanhatalo@helsinki.fi
[3]Amazon.com, gojav@amazon.com

Abstract

Bayesian optimization (BO) is a global optimization strategy designed to find the minimum of an expensive black-box function, by using a Gaussian process (GP) as a surrogate model for the objective. Although currently available acquisition functions address this goal with different degree of success, an over-exploration effect of the contour of the search space is typically observed. We propose a method to incorporate this knowledge into the searching process by adding virtual derivative observations in the GP at the borders of the search space. The method is applicable with any acquisition function, it is easy to use and consistently reduces the number of evaluations required to optimize the objective irrespective of the acquisition used. We illustrate the benefits of our approach in an extensive experimental comparison.

1 Introduction

Global optimization is a common problem in a very broad range of applications. Formally, it is defined as finding $\mathbf{x}_{\text{min}} \in \mathcal{X} \subset \mathcal{R}^d$ such that $\mathbf{x}_{\text{min}} = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$, where $\mathcal{X}$ is generally considered to be a bounded set. In this work, we focus on cases in which f is a black-box function whose explicit form is unknown and that it is expensive to evaluate. This implies that we need to find the $\mathbf{x}_{\text{min}}$ with a finite, typically, small number of evaluations.

Bayesian optimization of black-box functions using Gaussian Processes (GPs) as surrogate priors has become popular in recent years (see, e.g., review by Shahriari et al., 2016). A common problem that has not been systematically studied in the BO literature is the tendency of most acquisition strategies to over-explore the boundary of the function domain $\mathcal{X}$. This issue is not relevant if the global minimum may lie on the border of the search space but in most cases this is not the case. This effect has also been observed in the active learning literature (Krause and Guestrin, 2007). The unwanted boundary over-exploration effect of the regular BO is illustrated in Figure 1. In this paper we propose a new approach to correct the *boundary over-exploration effect* of most acquisitions.

2 Background and problem set-up

2.1 Standard GP surrogate for modeling of f

At iteration $n+1$, we assume that we have evaluated the objective function n times providing us the data $\mathbf{D} = \{y^{(i)}, \mathbf{x}^{(i)}\}_{i=1}^{n}$ where $y^{(i)}$ is, the possibly noisy, function evaluation at input location $\mathbf{x}^{(i)}$. A GP prior is directly specified on the latent function with prior assumptions encoded in the covariance function $k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, which specifies the covariance of two latent function values $f(\mathbf{x}^{(1)})$ and $f(\mathbf{x}^{(2)})$. A zero mean Gaussian process prior $p(\mathbf{f}) = \text{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$, is chosen, where $\mathbf{K}$ is

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

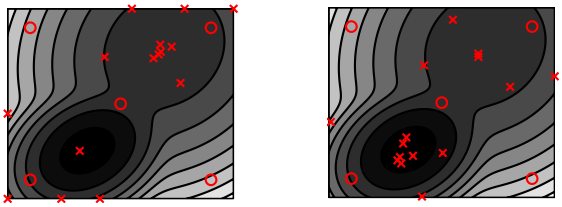

Figure 1: Sequence of 15 evaluations when optimizing a combination of Gaussians (darker colors represent lower function values) with standard BO on left and the proposal of this work on right. The five red circles are the points used to initialise the GP. The 15 red crosses are the acquisitions. With the new proposal, less evaluations are spent in the boundary, and more points are collected around the global optimum.

a covariance matrix between n latent values $\mathbf{f}$ at input used for training, $\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$, s.t. $\mathbf{K}_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

In regression, n noisy observations $\mathbf{y}$ and o latent function values $\mathbf{f}_*$ at the test inputs $\mathbf{X}_*$ are assumed to have Gaussian relationship. With the noise variance σ^2, the covariance between the latent values at the training and test inputs $\mathbf{K}_*$ and the covariance matrix of the latent values at the test inputs $\mathbf{K}_{**}$, the joint distribution of the observations and latent values at the test inputs is

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2\mathbf{I} & \mathbf{K}_*^{\mathrm{T}} \\ \mathbf{K}_* & \mathbf{K}_{**} \end{bmatrix}\right). \tag{1}$$

Using the Gaussian conditioning rule, the predictive distribution becomes $\mathbf{f}_*|\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$ with $\boldsymbol{\mu}_* = \mathbf{K}_*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$, and $\boldsymbol{\Sigma}_* = \mathbf{K}_{**} - \mathbf{K}_*(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_*^{\mathrm{T}}$.

2.2 Acquisition policies

Among others, some very well established acquisition functions are available. The *expected Improvement* (EI) maximizes the expected gain over the current best (Jones et al., 1998). The *lower confidence bound* (LCB) minimizes the regret over the optimization area (Srinivas et al., 2010) E. Brochu, V. M. Cora , and N. de Freitas (2010). The *probability of improvement* (MPI), selects the next point where probability of improving over the current best is the highest (Kushner, 1964).

2.3 Incorporating partial derivative observations in the loop

Since the differentiation is a linear operator, the partial derivative of a Gaussian process remains a Gaussian process (Solak et al., 2003; Rasmussen and Williams, 2006). Using this property, covariance matrices in equations of Section 2.1 can be extended to include partial derivatives either as observations or as values to be predicted. Following Riihimäki and Vehtari (2010), the probability of observing partial derivative of direction d, m_d, is modelled using probit likelihood with a control parameter ν (see details in Riihimäki and Vehtari, 2010) Assuming conditional independence given the latent derivative values, the likelihood of all the derivative values becomes a product of probit likelihoods. Now the joint posterior for the latent values and the latent value derivatives can be derived from the Bayes' rule. However, since there are probit components, the full posterior is analytically intractable and some approximation method must be used. Following Riihimäki and Vehtari (2010), we use expectation propagation (EP) for fast and accurate approximative inference.

3 Bayesian optimization with virtual derivative sign observations

Just like in the regular BO with GP prior presented in the Section 2, the objective function is given a GP prior which is updated according to the objective function evaluations so far. The next evaluation point is the acquisition function maximum, but if it is closer than threshold ϵ_b to the border of the search space, the point is projected to the border and a virtual derivative observation is placed to that point instead. After having added this virtual observation, the GP prior is updated and new proposal for the next acquisition is computed. Algorithm 1 contains pseudo code for the proposed method.

For more robustness, the following modifications can be used. Before placing a virtual derivative observation on the border, it can be checked whether or not the existing data supports the virtual

Algorithm 1 Pseudo code of the new BO method. The inputs are the acquisition a, the *stopping criterion* and the GP model. Note that the algorithm reduces to standard BO when lines 4-6 are removed.

1: **while** *stopping criterion* is False **do**
2: Fit GP to the available dataset $\mathbf{X}$, $\mathbf{y}$.
3: Optimise acquisition function, a, to find select new location $\mathbf{x}$ to evaluate.
4: **if** $\mathbf{x}$ is close to the edge **then**
5: Augment $\mathbf{X}$ with a virtual derivative sign observation at $\tilde{\mathbf{x}}$.
6: **else**
7: Augment $\mathbf{X}$ with $\mathbf{x}$ and evaluate g at $\mathbf{x}$.

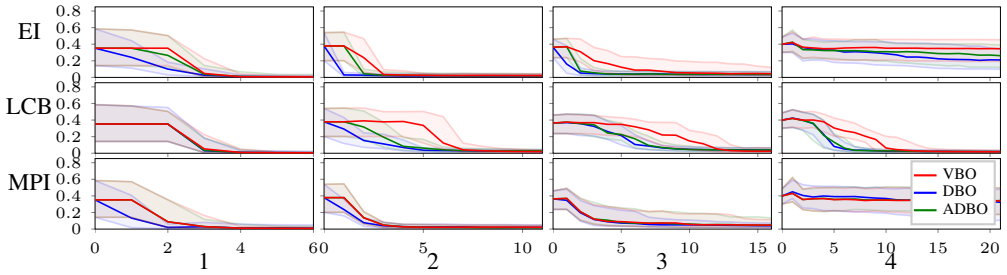

Figure 2: Each of the plot grids row illustrates results for different acquisition functions and each column illustrates functions of different dimension for case study 2.

gradient sign observation to be added to the model. If the energy of the assumed derivative sign is smaller than of opposite direction, it is reasonable to add the virtual observation. If there are minima on border, acquisitions might be proposed to locations where already exists virtual observations. If this happens, it is reasonable to remove the virtual observation before adding the new acquisition.

4 Experiments

The proposed Bayesian optimization algorithm was implemented in GPy toolbox [1]. In all four case studies to be covered later, we use zero mean GP prior for regular observations and probit likelihood with scaling parameter $\nu = 10^{-6}$ for the virtual derivative observations. In addition to this, we use the squared exponential covariance function. Three BO algorithms are used in the case studies. Standard BO algorithm (referred as vanilla BO, VBO), BO algorithm with virtual derivative sign observations (referred as derivative BO, DBO) and adaptive version of DBO (referred as ADBO). For the last two of these, virtual derivative sign observations are added if the next proposed point is within 1% of the length of the edge of the search space to any border. For ADBO, old virtual derivative observations are removed before adding regular observation if the euclidean distance between the points is less than 1% of the length of the edge of the search space.

4.1 Case study 1: A simple example function

The algorithm is used to illustrate the unwanted boundary over-exploration effect of the regular BO. To show this, simple function consisting of two Gaussian components is optimized with VBO and DBO using LCB as an acquisition function. The function and 15 first acquisitions are visualized in Figure 1, in the introduction.

4.2 Case Study 2: Random multivariate normal distribution functions

The algorithms are used to find the minimum of different d-dimensional negative multivariate normal distribution (MND) functions, for which the means are not located at the borders of the search space. To simulate real life observations, random noise $\epsilon \sim N(0, s)$ is added to the observations $y(\mathbf{x}) = g(\mathbf{x}) + \epsilon$, for some fixed s.

We drew 100 such functions per dimension in up to 4 dimensions, for three different noises $s = \{0, 0.05, 0.1\}$. For all these 400 functions, we ran the three BO algorithms for all three acquisition functions until the algorithms converged to the minimum. 25, 50, and 75 percentiles of found minimum values of these functions as a function of iterations are illustrated in Figure 2 for with

[1] Toolbox available at: `https://sheffieldml.github.io/GPy/`

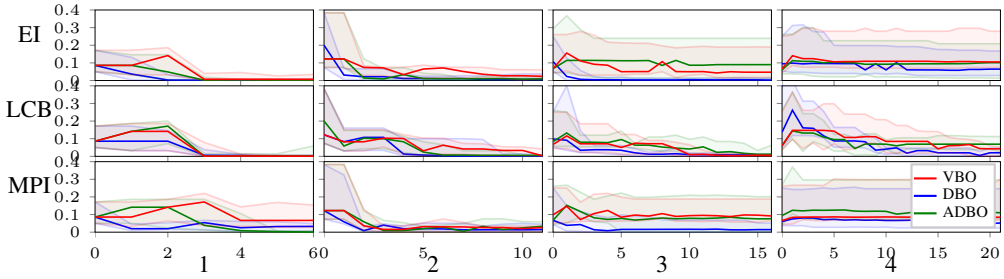

Figure 3: Same as in Figure 2, but with functions from Sigopt-library.

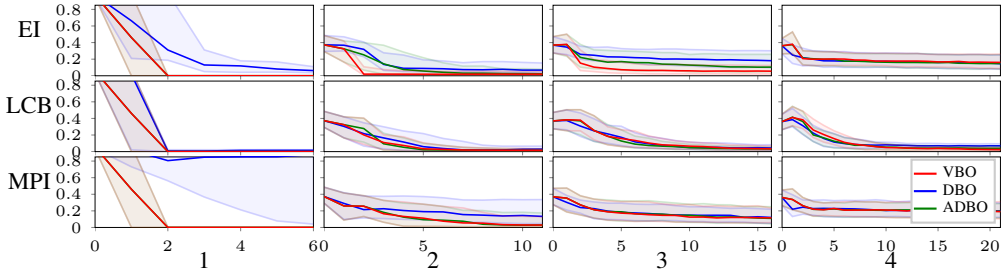

Figure 4: Same as in Figure 2, but with MND-functions that have local minimum on the edge of the search space.

$s = 0.1$. Results for other noise levels can be found from supplementary material. The results show that in all dimensions for all acquisition functions, performances of DBO and ADBO are better than or equal to the performance of VBO. Particularly with acquisition functions EI and LCB, DBO converges to the global minimum significantly faster than VBO, especially if the function is corrupted by noise. The variance of the optimization performance between different optimization runs is notably smaller for DBO and ADBO than for VBO.

4.3 Case study 3: Sigopt function library

Dewancker et al. (2016) introduced a benchmark function library called Sigopt for evaluating BO algorithms. When restricting the functions to at maximum 4 dimension and taking into account only unimodal, non-discrete functions with no global border minima or large plateaus, the library outputs 48 functions. The results are illustrated in Figure 3 for $s = 0.1$. Results for other noise levels can be found from supplementary material. The results are similar as for MND functions. DBO and ADBO still perform better than VBO, especially with LCB and EI. There is not as much difference between ADBO and VBO as before. Similarly as before, the variance of the optimization performance between different optimization runs is notably smaller for DBO and ADBO than for VBO.

4.4 Case study 4: Simple Gaussian functions with minima on border

The algorithms are used to find minimum of similar MND functions as in Section 4.2, with the difference that the global minima of each function is on the border of the search space. The results are illustrated in Figure 4 for $s = 0.1$. Results for other noise levels can be found from supplementary material. As expected, the results show that DBO does not perform as well as VBO and ADBO. The performances of ADBO and VBO are very similar for all noise levels and dimensions for LCB and MPI, for EI it performs similar as DBO.

5 Conclusions

We have presented here a Bayesian optimization algorithm which utilizes qualitative prior information concerning the objective function on the borders. Typical uses of Bayesian optimization concern expensive functions and in many applications qualitative knowledge of the generic properties of the function are known prior to optimization. The proposed BO method has proved to significantly improve the optimization speed and the found minimum when comparing the average performance to the performance of the standard BO algorithm without virtual derivative sign observations.

References

Dewancker, I., McCourt, M., Clark, S., Hayes, P., Johnson, A., and Ke, G. (2016). A stratified analysis of Bayesian optimization methods. *arXiv preprint arXiv:1603.09441*.

E. Brochu, V. M. Cora , and N. de Freitas (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492.

Krause, A. and Guestrin, C. (2007). Nonmyopic active learning of Gaussian processes: an exploration-exploitation approach. In *Proceedings of the 24th International Conference on Machine Learning*, volume 227 of *ACM International Conference Proceeding Series*, pages 449–456. ACM.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106.

Rasmussen, C. E. and Williams, C. K. I. (2006). Gaussian processes for machine learning. *The MIT Press*, 2(3):4.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 645–652.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

Solak, E., Murray, S. R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E. (2003). Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems*, pages 1033–1040.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. *Proceedings of the 27th International Conference on Machine Learning*.