Bayesian Optimization with Monotonicity Information

Cheng Li, Santu Rana, Sunil Gupta, Vu Nguyen, Svetha Venkatesh
Centre of Pattern Recognition and Data Analytic (PraDa), Deakin University
Email: cheng.l@deakin.edu.au

Abstract

Bayesian optimization (BO) has been demonstrated to be an efficient tool to globally optimize an expensive black-box function. Currently, however only a few works have explored the use of domain knowledge in BO to gain further efficiency. In this paper we discuss a particular form of prior information - the monotonicity of the underlying function with respect to one or more certain variables. Given the monotonicity information, we first detect the monotonic direction such as increasing or decreasing at each iteration. We then incorporate the detected monotonic direction into our proposed BO algorithm. We show the utility of our algorithm in target value optimization problems. Through the simulation we demonstrate the correctness of the proposed algorithm in discovering the monotonic direction. We also demonstrate the superiority of our algorithm in a real-world experimental optimization for short polymer fiber with the target geometric properties.

1 Introduction

Bayesian optimization (BO) has attracted significant research interests recently due to its efficiency in global optimization for black-box functions [10, 1, 7, 4]. However, only a few work has explored the use of prior knowledge about the underlying function in BO to further improve its efficiency. There are different types of prior knowledge about a function such as monotonicity [9], U-shape and S-shape [2] and quasiconvexity [3]. These priors have been used to improve function modeling. We discuss one particular form of prior knowledge - the monotonicity of a function with respect to one or more certain variables. In real applications experts sometimes have the prior knowledge that experiment result is monotonic with respect to one or more experimental parameters. For example, in short polymer fiber production the experts believe in advance that the fiber length is monotonically decreasing with the butanol speed [5]. To achieve a fiber with the target length, the experimenter often manually adjusts the experiment parameters based on the monotonicity information. Motivated by this, we propose to incorporate the monotonicity into BO to accelerate experimental design for the targeted product.

We formulate the target value optimization as the problem of minimizing the difference between the target value and the underlying function. Mathematically the objective is

$$\boldsymbol{x}^* = \text{argmin}_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x}) \triangleq \text{argmin}_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x}) - f_T|$$

where $f(\boldsymbol{x})$ is the underlying function and f_T is the target value. We can employ the standard BO approach to minimize $g(\boldsymbol{x})$. It first uses Gaussian process (GP) to model $g(\boldsymbol{x})$ and then constructs the acquisition function which is cheap to maximize to query the next point of $f(\boldsymbol{x})$ and proceeds this repeatedly. However, it is not clear how the monotonicity about $f(\boldsymbol{x})$ can be encoded into BO to facilitate the minimization of $g(\boldsymbol{x})$. Furthermore, if the monotonic direction of $f(\boldsymbol{x})$ is unknown, how can we detect and decide its monotonic direction such as increasing or decreasing before using the information?

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

To answer the first question, we propose a novel algorithm to incorporate the monotonicity into the BO. Suppose that we have the prior knowledge that $f(\boldsymbol{x})$ is monotonically increasing or decreasing with respect to the specified variables. We use GP to model $f(\boldsymbol{x})$ to make sure the mean function to be monotonic in these dimensions. It is achieved by following the work in [9] along with the positive or negative partial derivative signs in the whole search space. We then sample virtual observations from the GP of $f(\boldsymbol{x})$, which in turn will be combined with the actual observations to construct GP for $g(\boldsymbol{x})$. BO can finally be applied to the optimization of $g(\boldsymbol{x})$. In this way, we can transfer the monotonicity of $f(\boldsymbol{x})$ to $g(\boldsymbol{x})$ in its usable form.

To answer the second question, we use Bayesian model selection. We compute the leave-out-one (loo) predictive likelihood [11] for two monotonicity hypotheses on $f(\boldsymbol{x})$: monotonically decreasing and monotonically increasing. The one with the highest loo predictive likelihood is then selected to decide the monotonic direction of $f(\boldsymbol{x})$. In practice, at each iteration we decide the monotonic direction of the underlying function $f(\boldsymbol{x})$. We then utilize the positive or negative derivative signs corresponding to monotonically increasing and monotonically decreasing into our proposed algorithm. In short, our main contributions are:

- an algorithm to detect the monotonic detection of the underlying function for BO;
- the proposal of a novel BO algorithm to incorporate the monotonicity of the underlying function to optimize towards a target value;
- the validation of our proposed algorithm through both simulation and the experimental design of short polymer fiber with the target length.

2 The Proposed Algorithm

We know that the first order derivative signs of a monotonic function are always positive or negative. Riihimäki and Vehtari [9] developed an elegant framework to incorporate derivative signs into Gaussian process. Following this work, we derive the posterior GP with derivative signs. Then we propose a novel algorithm to encode the monotonicity of $f(\boldsymbol{x})$ in Bayesian optimization. Finally we detect the monotonic direction of $f(\boldsymbol{x})$ in the specified variables including monotonically increasing and monotonically decreasing based on the current observations.

2.1 Gaussian process with derivative signs

Since the GP is a linear operator, the derivative of Gaussian process is still a Gaussian process [8]. Therefore, it is flexible to incorporate derivative values into GP for prediction. Let $\{\boldsymbol{x}_i, y_i\}_{i=1}^t$ be observations and the ith sample $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{noise})$, as the noisy observation of $f(\boldsymbol{x})$ at $\boldsymbol{x}_i$. We denote $X = \{\boldsymbol{x}_i\}_{i=1}^t$ and $\boldsymbol{y} = \{y_i\}_1^t$. Let $X_s = \{\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2}, \cdots, \boldsymbol{x}_{s_m}\}$ be the locations of virtual derivative observations and $\boldsymbol{s} = \{s_1, s_2, \cdots, s_m\}$ be partial derivative signs for the variables $\boldsymbol{d}$. The latent function value and the partial derivative value for the variables $\boldsymbol{d}$ are denoted as $\boldsymbol{f}$ and $\boldsymbol{f}^{'}$ respectively.

Riihimäki and Vehtari [9] have employed a probit function to build the link derivative sign $\boldsymbol{s}$ and derivative value $\boldsymbol{f}^{'}$

$$p(\boldsymbol{s}|\boldsymbol{f}^{'}) = \prod_{i=1}^{m} \Phi\left(\frac{s_i \partial f^{(i)}}{\partial x_d^{(i)}} \frac{1}{v}\right) \tag{1}$$

where $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(x \mid 0, 1)dx$ and the steepness v controls the slope of monotonicity. Riihimäki and Vehtari [9] use expectation propagation to approximate Eq.(1). We refer the readers to go through the detail inference for GP with derivative signs in [9].For a new point, the predictive mean and variance has the same form in GP with derivative signs as the standard GP [8]. In our experiments, we empirically place virtual derivative observations in a grid.

2.2 Bayesian optimization with monotonicity

We propose a new algorithm to encode the monotonicity of $f(\boldsymbol{x})$ into the BO of $g(\boldsymbol{x})$. In this algorithm we first make the mean function of $f(\boldsymbol{x})$ to be monotonic in Gaussian process and then we sample points from this GP which in turn is combined with existing observations to build a new GP

Algorithm 1 Bayesian optimization with monotonicity Information

Input: observations $\mathcal{D}_{1:t} = \{\boldsymbol{x_i}, y_i\}_{i=1}^t$ on $f(\boldsymbol{x})$, the target value f_T, the specified variables $\boldsymbol{d}$ for monotonic direction detection

obtain the observations $\mathcal{G} = \{\boldsymbol{x}_i, |y_i - f_T|\}_{i=1}^t$ of $g(\boldsymbol{x})$;
for $t = 1, 2, \cdots$ **do**
  perform monotonic direction detection on $f(\boldsymbol{x})$ (sec 2.3);
  build Gaussian process with derivative signs (positive or negative) on $f(\boldsymbol{x})$ (sec 2.1);
  sample the virtual observations $\mathcal{V}$ from the monotonic GP above (sec 2.2);
  build Gaussian process on $g(\boldsymbol{x})$ using $\mathcal{V}$ and $\mathcal{G}$ (Eq.(2) and (3)) (sec 2.2);
  sample the next point $\boldsymbol{x}_{t+1} \leftarrow \text{argmax}_{\boldsymbol{x}_{t+1} \in \mathcal{X}} a(\boldsymbol{x} \mid \mathcal{G}, \mathcal{V})$;
  evaluate the function $y_{t+1} = f(\boldsymbol{x}_{t+1}) + \varepsilon$;
  augment the data $\mathcal{D}_{1:t+1} = \{\mathcal{D}_{1:t}, \{\boldsymbol{x}_{t+1}, y_{t+1}\}\}$;
end for

model on $g(\boldsymbol{x})$. By this way we make full use of the monotonicity of $f(\boldsymbol{x})$ and transfer this critical knowledge to $g(\boldsymbol{x})$ through the sampled points. To be specific, we model the function $f(\boldsymbol{x})$ using monotonic GP described in section 2.1. We then randomly sample J points $X_v = \{\boldsymbol{x}_p^v\}_{j=1}^J$ from the monotonic GP. We denote the sampled set $\mathcal{V} = \{\boldsymbol{x}_p^v, \mu_f(\boldsymbol{x}_p^v), \sigma_f^2(\boldsymbol{x}_p^v)\}_{j=1}^J$ with the mean and variance. Combing sampled points and existing observations $\{\boldsymbol{x}_i, |y_i - f_T|\}_{i=1}^t$, we can build a new GP on $g(\boldsymbol{x})$ and then perform Bayesian optimization. The mean and variance for a new point $\boldsymbol{x}_{t+1}$ in this GP are

$$\mu_g(\boldsymbol{x}_{t+1}) = \mathbf{k}^T K^{-1} [\boldsymbol{\mu}_g(X_v); \boldsymbol{\mu}_g(X)] \tag{2}$$

$$\sigma_g^2(\boldsymbol{x}_{t+1}) = 1 - \mathbf{k}^T K^{-1} \mathbf{k} \tag{3}$$

where $\mathbf{k} = [k(\boldsymbol{x}_{t+1}, \boldsymbol{x}_1^v) \cdots k(\boldsymbol{x}_{t+1}, \boldsymbol{x}_J^v)\, k(\boldsymbol{x}_{t+1}, \boldsymbol{x}_1) \cdots k(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t)]$, $\boldsymbol{\mu}_g(X_v) = |\boldsymbol{\mu}_f(X_v) - f_T|$ and $\boldsymbol{\mu}_g(X) = |\boldsymbol{y} - f_T|$,

$$K = \begin{bmatrix} K_{VV} & K_{VX} \\ K_{XV} & K_{XX} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\sigma}_f^2(X_v) & \mathbf{0} \\ \mathbf{0} & \sigma_{noise}^2 \end{bmatrix} \mathbf{I}$$

where K_{VV} are the self-covariance matrix of X_v and K_{XV} is the covariance matrix between X and X_v.

2.3 Monotonic direction detection

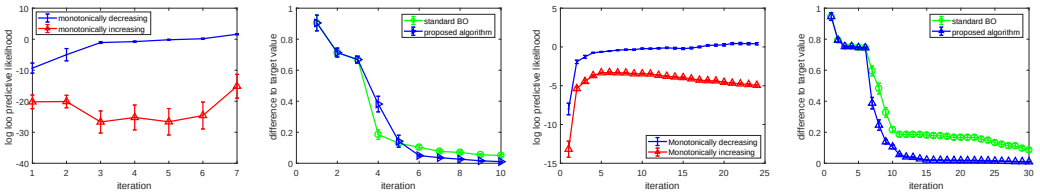
In experimental designs, experts sometimes have the prior monotonicity about the experiment result with respect to one or more variables. However they might not be confident about the monotonic direction such as increasing or decreasing. We analytically decide the monotonic direction based on the current observations so that we can flexibly incorporate the correct derivative signs into our algorithm in sec 2.2.

Suppose we are given a set of observed data $\mathcal{D} = \{X, \boldsymbol{y}\}$. Given monotonicity hypotheses such as monotonically increasing and monotonically decreasing with respect to one ore more specified variables, we can train Gaussian process models corresponding different hyperparameters $\boldsymbol{\theta}_i$. The first object of interest for prediction assessment in Bayesian model is the leave-one-out (Loo) predictive likelihood [6]. It represents the likelihood of the left-out point given a model trained on other points. We compute it as

$$Loo = \frac{1}{t} \sum_{i=1}^{t} p(y_i \mid \boldsymbol{x}_i, \mathcal{D}_{-i}) \tag{4}$$

$$p(y_i \mid \boldsymbol{x}_i, \mathcal{D}_{-i}) = \int p(y_i \mid \boldsymbol{x}_i, \theta) p(\theta \mid \mathcal{D}_{-i}) d\theta \tag{5}$$

where $\mathcal{D}_{-i}$ is all other observations except $(\boldsymbol{x}_i, y_i)$. In Gaussian process, Eq.(5) is a Gaussian distribution and thus Eq.(4) is tractable. At each iteration, we compare the leave-out-one predictive likelihood of two hypotheses: monotonically increasing and monotonically decreasing and choose one with the highest likelihood value. We then adopt the algorithm in sec 2.2 to incorporate positive or negative derivative signs of $f(\boldsymbol{x})$ to perform BO. The algorithm we propose is presented in Alg 1.



(a) 2D function f_1. (b) 5D function f_2.

Figure 1: The simulation results for benchmark functions. Left in (a) and (b): The comparison of the log leave-out-one (loo) predictive likelihood between monotone increasing and monotone decreasing. Right in (a) and (b): The comparison of the standard BO and the proposed algorithm. The vertical axis represents the difference to the target value.

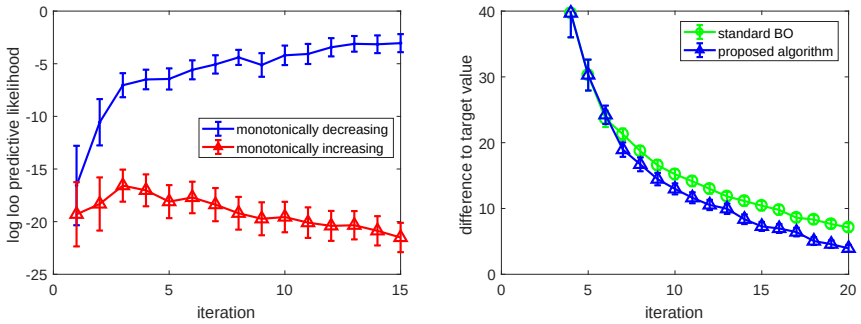


Figure 2: The real experiment for optimizing the SPF with the target length $70\mu m$. Left: The comparison of the log leave-out-one (loo) predictive likelihood between monotone increasing and monotone decreasing. Right: The comparison of the standard BO and the proposed algorithm. The vertical axis represents the difference to the target value.

3 Experiments

We compare the proposed algorithm to the standard BO on the applications of benchmark function optimization as well as optimization of short polymer fibers with targeted length. Since the parameter ν in Eq.(1) reflects the slope of the monotonic function, we empirically set $\nu = 0.01$ in all experiments. For hyperparameters in all GP, we automatically estimated them at each iteration by maximizing the marginal likelihood. We run all experiments for 50 times and report the average mean and the standard error. We optimize benchmark functions which is monotonically decreasing with respect to some variables. The benchmark functions we use are:

(a) 2D function: $f_1(\boldsymbol{x}) = \frac{1}{20}(x_1 - 5)^2 + \frac{1}{20}(x_2 - 5)^2$, $f_T = 1.5$, $\boldsymbol{x} \in [0, 5]$;

(b) 5D function: $f_2(\boldsymbol{x}) = \frac{1}{30}(x_1 - 4)^2 + \frac{1}{30}(x_2 - 4)^2 + \mathcal{GN}(x_{3:5}|\mathbf{0}, \mathbf{1})$, $f_T = 1.5$, $\boldsymbol{x} \in [-2, 4]$, where $\mathcal{GN}(x_{3:5}|0, 1)$ is a un-normalized Gaussian PDF for $x_3 \sim x_5$;

The $D + 1$ initial points are randomly sampled from synthetic functions. For both the f_1 and f_2, we test our algorithm by leveraging the information that the function is monotonic with respect to x_1. We first compute the log loo predictive likelihood of two hypotheses consisting of monotone increasing and decreasing based on the current observations and then we run BO with the detected monotonic direction. The experiment results are shown in Fig 1. For $f_1$we detected that the function is monotonically decreasing and subsequently the proposed algorithm can approach the target more quickly than the standard BO. Similarly for f_2, our algorithm outperform the standard BO significantly.

We also test our algorithm on the real-world application of optimizing the short polymer fiber (SPF) with the target length. To simplify the problem, we optimize five parameters including channel width, butanol speed, constriction angle, polymer concentration and device position to produce the desirable fiber [5]. Material experts have provided us the information that the fiber length is monotonic wrt the

butanol speed. Thus the goal of our task is to decide the correct monotonic direction and leverage the decision to facilitate the optimization of the fiber with the target length. We set the target length of the fiber as 70μm in experiments and used 5 random experiments intiially. The log loo predictive likelihood demonstrated in Fig 2 shows that we detected that the fiber length is monotonically decreasing with the butanol speed and the proposed algorithm can approach the target faster and thus reducing the number of real experiments.

4 Conclusion

We have proposed a completed algorithm for monotonic direction detection and the incorporation of the monotonicity information about the underlying function into the BO framework. The experiment results have shown that the proposed algorithm significantly outperforms the standard BO performance. Regarding the work in this paper we will seek for a smart way to automatically detect the trend of the function without any assumption so that different BO strategies can switch freely between each other. More broadly we have envisaged the benefit of the use of monotonicity information in BO and therefore exploring the use of other types of prior knowledge in BO is a promising direction.

References

[1] Eric Brochu, Vlad M. Cora, and Nando De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[2] Taeryon Choi and Peter J. Lenk. Bayesian analysis of shape-restricted functions using gaussian process priors. *Statistica Sinica*, 27(1):43–69, 2017.

[3] Michael Jauch and Víctor Peña. Bayesian optimization with shape constraints. In *Advances in Neural Information Processing Systems 2017 Workshop*, 2016.

[4] Cheng Li, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, and Aistair Shilton. High dimensional bayesian optimization using dropout. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2096–2102, 2017.

[5] Cheng Li, David Rubín de Celis Leal, Santu Rana, Sunil Gupta, Alessandra Sutti, Stewart Greenhill, Teo Slezak, Murray Height, and Svetha Venkatesh. Rapid bayesian optimisation for synthesis of short polymer fiber materials. *Scientific Reports*, 7, 2017.

[6] Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, May 2017.

[7] Santu Rana, Cheng Li, Sunil Gupta, Vu Nguyen, and Svetha Venkatesh. High dimensional Bayesian optimization with elastic Gaussian process. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2883–2891, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[8] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[9] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 645–652, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[10] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *NIPS*, pages 2951–2959, 2012.

[11] Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther. Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *Journal of Machine Learning Research*, 17:103:1–103:38, 2016.