Bayesian Optimization of Unimodal Functions

Michael Riis Andersen, Eero Siivola, and Aki Vehtari
Helsinki Institute for Information Technology, HIIT
Department of Computer Science, Aalto University, Finland

Abstract

Bayesian Optimization (BO) is a global optimization strategy designed to find the minimum of a black-box function by using a Gaussian process (GP) as a surrogate model for the function to be optimized. In this work, we study learning and optimization of unimodal functions using Bayesian optimization. We propose a hierarchical model for unimodal functions based on Gaussian processes with virtual derivative observations. We demonstrate that taking such structural prior information into account can decrease the number of function evaluations significantly and improve data efficiency.

1 Introduction

Bayesian optimization has proved itself a valuable tool for global optimization of expensive objective functions. The objective functions are in general assumed to be black-box functions, whose analytical form is unknown, but expensive to evaluate. However, if there is more information available, this information should be taken into account in order to minimize the number of required function evaluations. We focus on the case where the unknown objective functions are assumed to be unimodal. Unimodality is a natural assumption in many applications, including modelling dose-response relationships [1, 2], Approximate Bayesian Computation (ABC) [3], and density estimation [4].

A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is said to be unimodal if there exist a point $c_0 \in \mathbb{R}$ such that $f(x)$ is monotonically decreasing for $x < c_0$ and monotonically increasing for $x > c_0$. Consequently, the function f has a single minima at $x = c_0$. In this work, we study learning and optimization of unimodal functions using the Bayesian optimization framework [5]. We will assume that f is unimodal with a single minimum rather than a single maximum without loss of generality.

We propose a hierarchical model for unimodal functions based on Gaussian processes [6], where the unimodality assumption is encoded using virtual derivative observations [7]. The contribution of this paper is two-fold. First, we describe a Gaussian process-based model for unimodal functions (Section 2), which can be applied to unimodal regression problems in general. Second, we demonstrate that taking advantage of such structural prior knowledge can lead to improved data efficiency in the Bayesian optimization setting (Section 3).

2 Unimodal Regression using Gaussian processes

In this section, we will first briefly review Gaussian processes (GPs) and Gaussian processes with derivative information. Finally, we describe our proposed approach to unimodal regression based on GPs .

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

2.1 Gaussian Process Regression with Derivative Information

Gaussian processes are a flexible nonparametric family of distributions over functions [6]. Consider a dataset $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$, where y_n is a noisy observation of f at x_n, i.e. $y_n = f(x_n) + \epsilon_n$. We assume a Gaussian process prior distribution for $f \sim \mathcal{GP}\left(\mu(\cdot), k(\cdot,\cdot)\right)$, where $\mu : \mathbb{R} \rightarrow \mathbb{R}$ and $k : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ are the mean and covariance functions, respectively. Assuming an isotropic Gaussian observation model, the joint distribution becomes

$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{f}, \sigma^2 \boldsymbol{I}\right)\mathcal{N}\left(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{K}\right), \tag{1}$$

where $\boldsymbol{f}_n = f(x_n)$, $\boldsymbol{\mu}_n = \mu(x_n)$, $\boldsymbol{K}_{mn} = k(x_m, x_n)$, and $\sigma^2 > 0$ is the noise variance. As the joint distribution is Gaussian, any conditional or marginal distribution is readily available in closed form. Similarly, the predictive distribution for a test point x^* is also readily obtained by extending the joint distribution to include $f^* = f(x^*)$ and then computing the marginal posterior $p(f^*|\mathcal{D})$.

The derivative of a (differentiable) Gaussian process realization is also characterized by a Gaussian process as differentiation is a linear operator [8]. The first and second moments of the derivative process can be obtained as partial derivatives of the moments of the process f [6]. By formulating a joint model of f and f', it is possible to include derivative information in Gaussian process models. Structural constraints such as monotonicity can be induced by introducing virtual observations of the sign of the derivative function [7].

2.2 Unimodal Gaussian Processes Regression

If f is a differentiable unimodal function with minima c_0, then it holds that $f'(x) < 0$ for $x < c_0$ and $f'(x) > 0$ for $x > c_0$. Based on this observation, we propose a hierarchical model that induces unimodality by introducing a set of virtual of derivative observations that satisfy these constraints.

Let $p(\boldsymbol{f}, \boldsymbol{f}')$ be the joint model of a Gaussian process f and its derivative f'. Further, suppose we observe the sign of f' when evaluated at a set of point in the input space $\{\tilde{x}_j\}_{j=1}^{J}$, i.e. $z_j = \text{sign}\left[f\left(\tilde{x}_j\right)\right] \in \{-1, 1\}$ for $j = 1 \dots J$. The joint model for a unimodal function f with observations $\mathbf{y}$ then becomes

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{f}', \boldsymbol{z}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, \boldsymbol{f}')\prod_{j=1}^{J} p(z_j|f'_j) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, \boldsymbol{f}')\prod_{j=1}^{J}\phi(\nu_f z_j f'_j), \tag{2}$$

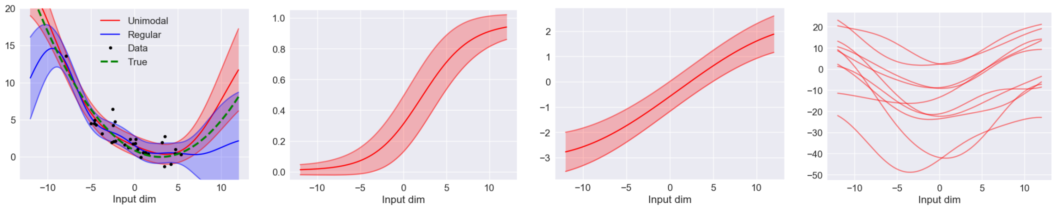
where ϕ is the cumulative distribution function of a standardized Gaussian random variable and $\nu_f > 0$ is a hyperparameter. The virtual likelihood term $\phi(\nu_f z_j f'_j)$ removes probability mass from functions that violate the constraint on the sign of the derivative at $\tilde{x_j}$.

However, observing z_j is usually not feasible in practice. Instead we propose to model z_j by taking advantage of the observation that $z_j = -1$ for $\tilde{x}_j < c_0$ and $z_j = 1$ for $\tilde{x}_j > c_0$. Thus, we can model the binary sign observations as $z_j | x_j \sim \text{Ber}\left(\phi\left(g(x_j)\right)\right)$, where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a non-decreasing function. For this, we use a Gaussian process with monotonicity constraints [7] as a prior distribution for g. Specifically, we introduce another Gaussian process g, its derivative g' and virtual observations of the sign of g', which gives rise to the following model for $\boldsymbol{z}$

$$p(\boldsymbol{z}, \boldsymbol{g}, \boldsymbol{g}') \propto \prod_{j=1}^{J}\text{Ber}\left(z_j|\phi\left(g_i\right)\right)p(\boldsymbol{g}, \boldsymbol{g}')\prod_{j=1}^{J}\phi\left(\nu_g g'_j\right), \tag{3}$$

where $p(\boldsymbol{g}, \boldsymbol{g}')$ is the joint Gaussian distribution of the g and g' evaluated at the points $\{\tilde{x}_j\}_{j=1}^{J}$ and the virtual likelihood terms $\phi\left(\nu_g g'_j\right)$ for $j = 1 \dots J$ are constraining the derivatives of g' to be positive and thereby forcing g to be non-decreasing. Finally, combining the models in eq. (2) and (3), and marginalizing with respect to $\boldsymbol{z}$ gives rise to the following joint model

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{f}', \boldsymbol{g}, \boldsymbol{g}') &\propto \sum_{\boldsymbol{z}} p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, \boldsymbol{f}')\prod_{j=1}^{J}\phi(\nu_f z_j f'_j)\text{Ber}\left(z_j|\phi\left(g_i\right)\right)p(\boldsymbol{g}, \boldsymbol{g}')\prod_{j=1}^{J}\phi\left(\nu_g g'_j\right) \qquad (4) \\ &\propto p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, \boldsymbol{f}')\prod_{j=1}^{J}\left[\phi(-\nu_f f'_j)\phi\left(-g_i\right) + \phi(\nu_f f'_j)\phi\left(g_i\right)\right]p(\boldsymbol{g}, \boldsymbol{g}')\prod_{j=1}^{J}\phi\left(\nu_g g'_j\right). \end{aligned}$$



(a) (b) (c) (d)

Figure 1: a) Posterior mean and standard deviation for f for a toy data set with $N = 30$ observations of the unimodal function $y_n = 0.1(x_n - 3)^2 + e_n$, where $x_n \sim \mathcal{N}\left(0, 3^2\right)$ and $e_n \sim \mathcal{N}\left(0, 3\right)$. b) Posterior distribution of positive derivatives of g, i.e. $p(z_j = 1|\mathcal{D})$. c) Posterior distribution of g. d) 10 realizations from the prior of f.

The marginal posterior $p(\boldsymbol{f}|\mathcal{D})$ of the model in eq. (4) can then be interpreted as a Gaussian process subject to a unimodality constraint. Similarly, the model can easily be extended to include $f^* = f(x^*)$ in order to make predictions for test point x^*. Thus, the model can be applied for probabilistic unimodal regression as well as Bayesian optimization of unimodal functions. The posterior distributions of interest, $p(\boldsymbol{f}|\mathcal{D})$ and $p(f^*|\mathcal{D})$, are intractable and we have to resort to approximate inference. We used the expectation propagation algorithm [9]. In this work, we have focused on the one dimensional case as a proof of concept. However, the model can naturally be extended to the D dimensions by introducing D Gaussian processes, where the i'th GP, $g_i : \mathbb{R}^D \rightarrow \mathbb{R}$, controls the sign of the i'th the partial derivative, $\frac{\partial}{\partial x_i} f$. This approach also provides the flexibility to only impose unimodality with respect to a subset of the dimensions, if desired.

3 Numerical experiments

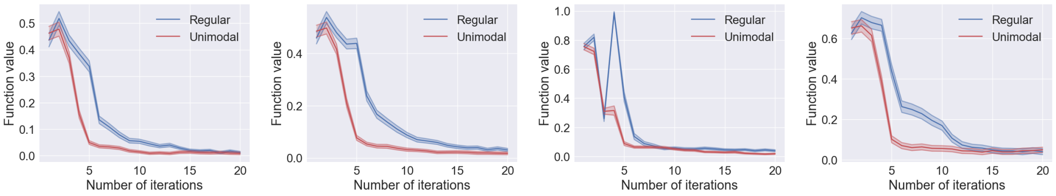
In this section, we describe three experiments that demonstrates the properties of the proposed method.

3.1 Unimodal Regression using Toy Data

The purpose of the first experiment is to demonstrate how the model can be applied to unimodal regression problems. We consider a toy data set with $N = 30$ noisy observations of the unimodal function $y_n = 0.1(x_n - 3)^2 + e_n$, where $x_n \sim \mathcal{N}\left(0, 3^2\right)$ and $e_n \sim \mathcal{N}\left(0, 3\right)$. We used $J = 20$ virtual observations evenly distributed in the interval $[-12, 12]$. We used a squared exponential function plus a constant as covariance function for f and a squared exponential covariance function for g. Both f and g were assumed to have zero mean. We used maximum a posteriori estimators for all kernel hyperparameters and for the noise variance. We imposed a half Student-t distribution for the prior variance of g and assumed uniform distributions for the remaining hyperparameters. Panel (a) in Figure 1 shows the posterior distribution for the proposed model and for a regular GP. Panel (b) and (c) show the posterior probability of the $z_j = 1$ and the posterior distribution for g, respectively. Panel (d) shows 10 realizations from the prior of f.

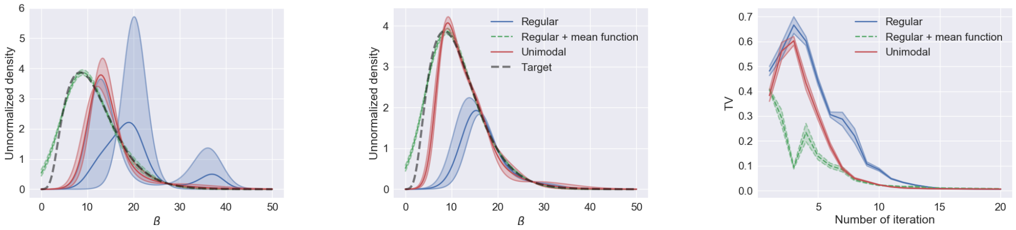
3.2 Simulation study

In the second experiment, we conducted a simulation study for Bayesian optimization of unimodal functions from four different classes of functions. The first three classes were the negative densities of univariate Gaussian, Student-t, and beta distributions and the fourth class was (scaled and translated) negative Tukey window functions. We sampled 200 functions from each class with random parameters. All functions were normalized to have both domain and image in the unit interval and all functions were corrupted with Gaussian noise of variance $\sigma^2 = 0.05^2$. Using three initial observations at $x \in \{0.2, 0.5, 0.8\}$, all functions were optimized using Bayesian optimization with expected improvement as acquisition function [5] for 20 iterations. We used the same kernels as described in the previous experiment, but now we used half Student-t priors for both the lengthscale and variance parameters of the squared exponential covariance function for f. Furthermore, we used lognormal priors for the prior variance and lengthscale of g and an inverse Gamma prior for the noise variance.



(a) Gaussian (b) Student-t (c) Beta (d) Tukey

Figure 2: Results from a simulation study of Bayesian optimization of unimodal functions from four different classes of functions: negative densities for Gaussian, student t, and beta distributions as well as translated and scaled negative Tukey windows. The results are averaged over 200 realizations for each class. The expected improvement acquisition function were used for all four classes.



(a) (b) (c)

Figure 3: Learning the marginal posterior density of β for the Bioassay data set using. a) Posterior distributions for the (unnormalized) density for after $N = 4$ iterations. b) Posterior distributions for the (unnormalized) density for after $N = 6$ iterations. c) Total variation as a function of number of iterations averaged across 100 runs.

Figure 2(a)-(d) shows the function values as a function of the number of iterations for each of the four function classes averaged across 200 realizations. It is seen that the unimodal GP significantly outperforms the regular GP for all four function classes.

3.3 Density estimation

GPs have been applied as surrogate density models for problems, where the exact posterior density is prohibitively expensive to evaluate [4]. In the final experiment, we demonstrate the benefit of the unimodal model for this application in a setting, where the true density is readily available to facilitate a quantitative evaluation of the proposed model. Consider the posterior distribution of a generalized linear model with binomial observations, $y_i|\theta_i \sim \text{Bin}\left(\theta_i, n_i\right)$, and a logit link function, $\text{logit}\left(\theta_i\right) = \alpha + \beta x_i$, with uniform prior distributions, $p(\alpha, \beta) \propto 1$. Using the unimodal GP, we model the negative logarithm of the marginal posterior density of β for a Bioassay data set [10]. We approximate the (unnormalized) density as $\hat{p}(\beta) \propto \mathbb{E}\left[\exp\left(-f(\beta)\right)\right]$, where the expectation is with respect to the posterior distribution of f. To learn the density function using as few evaluations as possible, we use the variance of $\hat{p}(\beta)$ as acquisition function. Starting from a single initial evaluation, we perform Bayesian optimization for 20 iterations and for every iteration we compute the total variation (TV) between the target density and the estimated density. We use a regular GP and a regular GP with a Gaussian mean function (Laplace approximation) as baselines. Figure 3(a) and (b) show the posterior for the (unnormalized) densities after $N = 4$ and $N = 6$ iterations, respectively. Panel (b) shows the total variation as a function of number of iterations averaged over 100 runs.

4 Summary

We proposed a probabilistic method for unimodal regression using Gaussian processes. We have demonstrated the method can be applied to learning and optimization of unimodal functions. In this work, we focused on univariate problems to provide proof of concept, but we will extend it to higher dimensions in future work.

References

[1] J Hardwick and Q F Stout. Optimizing a unimodal response function for binary variables. In *Optimum Design 2000*, Nonconvex Optimization and Its Applications, pages 195–210. Springer, Boston, MA, 2001.

[2] Q F Stout. Optimal algorithms for unimodal regression. *Ann Arbor*, 1001:48109–42122, 2000.

[3] M Sunnåker, A G Busetto, E Numminen, J Corander, M Foll, and C Dessimoz. Approximate Bayesian computation. *PLoS Comput. Biol.*, 9(1):e1002803, January 2013.

[4] C E Rasmussen, J M Bernardo, M J Bayarri, J O Berger, A P Dawid, D Heckerman, Afm Smith, and M West. Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals. In *Bayesian Statistics 7*, pages 651–659, 2003.

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

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

[7] J Riihimäki and A Vehtari. Gaussian processes with monotonicity information. *of the Thirteenth International Conference on . . .* , 2010.

[8] S Banerjee and A E Gelfand. On smoothness properties of spatial processes. *J. Multivar. Anal.*, 84(1):85–100, January 2003.

[9] T Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 362–369, San Francisco, CA, 2001. Morgan Kaufmann.

[10] A Gelman, J B Carlin, H S Stern, D B Dunson, A Vehtari, and D B Rubin. *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 3 edition edition, November 2013.