Bayesian Optimization

Nam-Anh Tran

1 Summary

The central theme of this review is Bayesian Optimization (BO). We will primarily focus on *Bayesian Optimization* by Frazier[1] while also considering *Optimization Under Unknown Constraints* by Gramacy et al.[2] The latter serves as an important extension of BO, particularly in addressing constraints.

The primary objective of the first study is to introduce how Bayesian optimization works, including Gaussian process regression and three standard acquisition functions: expected improvement (EI), entropy search (ES), and knowledge gradient (KG), but the focus is the first two functions. The paper also discussed more advanced techniques, including expensive-to-evaluate constraint functions, which were discussed in detail in the second paper. BO is a class of machine-learning-based optimization methods designed for functions with expensive evaluations or a lack of a known special structure. Thus, BO was utilized for black-box derivative-free global optimization.

BO involves two key components: a Bayesian statistical model for modelling the objective function and an acquisition function for deciding where to sample next. Following the evaluation of the objective, pertaining to an initial space-filling experimental design with points uniformly selected at random, the algorithm iteratively allocates the remainder of a budget of N function evaluations. The Gaussian process statistical model provides a posterior distribution to describe potential values at a candidate point. The posterior distribution is updated after we observe the evaluation of the new point. The acquisition function measures the value generated by evaluating the objective function at a new point based on the current posterior distribution. EI is the most commonly used acquisition function.

While EI only considers the posterior at the point sampled, the KG considers the posterior over the full domain of the objective function and how the sample will change that posterior. This is possible as it assess the improvement based on the mean function rather evaluation value of the points. KG provides a small performance benefit in the standard BO problem with noise-free evaluation[3] and substantial performance improvements in the problems with noise, multi-fidelity observations, derivative observations, the need to integrate environmental conditions, and other more exotic problem features.

Further, constraint BO (CBO) was briefly introduced in this paper. While the standard BO assumes the feasible set is simple and easy to assess membership, CBO adds functional constraints to the objective function to introduce an invalid domain where the points cannot be evaluated or accessed. Gramacy et al. introduced a new acquisition function, the Integrated Expected Conditional Improvement Algorithm (IECI), obtained by modifying EI so that IECI captures the improvement at each reference point in the feasible set conditional on the new potential point added yet evaluated. This idea is similar to KG, where we consider the stage where a new design point is supposed to be added to the design, but we have not evaluated it. The "integrated" part refers to the improvement over the whole feasible set; this step requires knowing the weights of the design points. The constrained set is, thus, reflected based on the weights, where those invalid points have zero weight. Unlike EI, IECI does not have a closed-form; thus, it is approximated using Monte Carlo.

2 Critical Review

BO is appealing because of its invasive application in various areas. Although the authors only mentioned its application in machine learning, specifically training hyper-parameters, BO can be used in optimal designs and dose finding in drug development, where the information about the structure of the objective function is vague. BO can also be used for functions with complex structures, i.e., the functional structure is known but differentiable, as it does not require calculating the derivative.

Gaussian processes are used in most work on BO, but this assumption may not be true for applications of interest. Thus, developing new statistical models for BO is still an open area. As the hyperparameters of Gaussian processes are considered in the model, BO quickly becomes more complicated and requires solving another optimization problem to estimate those hyperparameters, such as coefficients, scale and nugget parameters in the defined covariance matrix, if we employ empirical Bayes; on the other hand, the optimization algorithm will be expensively time-consuming if we employ fully Bayesian approach, where we have to simulate samples of the hyperparameters. Moreover, despite being faster, empirical Bayes suffers from the classic "flat-likelihood", i.e. the surface can be nearly flat, yielding multiple local maxima that all look equally good.

3 Application

As the first paper provided the the basic pseudo-code of BO, we will define a function that we optimize using BO and R built-in function (not BO) and compare the results. We define the following function

$$f(x) = -6 \times \phi(x|4, 1) - 7.5 \times \phi(x|7, 1) + 1.3,$$

where ϕ is the normal density function. We optimize the function using R `optimize` function. We then optimize the function using BO with EI acquisition function and compare the results. In this setting, the model only include the intercept at zero, and the covariance function is the inverse of the exponentiated squared Euclidean distance, i.e., $\exp\{-(x-x')^2\}$, $\forall x, x' \in \mathcal{X}$. Also, we do not consider the white noise, i.e. we assume that we can always obtain the true evaluations without noise.

Figure A1 show the objective function and samples generated using Gaussian process before we observe the data.

References

[1] Frazier PI. Bayesian optimization. In: *Recent Advances in Optimization and Modeling of Contemporary Problems*. Informs; 2018:255-278.

[2] Bernardo J, Bayarri M, Berger J, Dawid A, Heckerman D, Smith A, et al. Optimization under unknown constraints. *Bayesian Statistics*. 2011;9(9):229.

[3] Frazier P, Powell W, Dayanik S. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*. 2009;21(4):599-613.

Appendix

Figure A1: The objective function and samples generated using a multivariate normal distribution with mean zero and a covariance matrix calculated based on the the exponentiated squared Euclidean distance. The vertical and horizontal lines are the minimizer and minimum value obtained using the built-in R function.

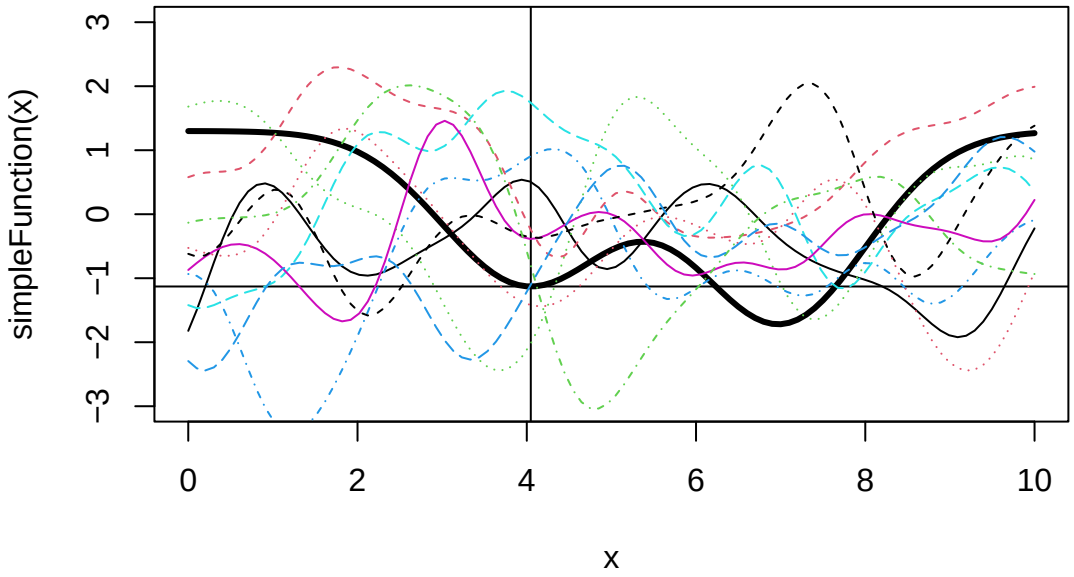

Figure A2: the plot shows the reduction of variance at evaluated locations.

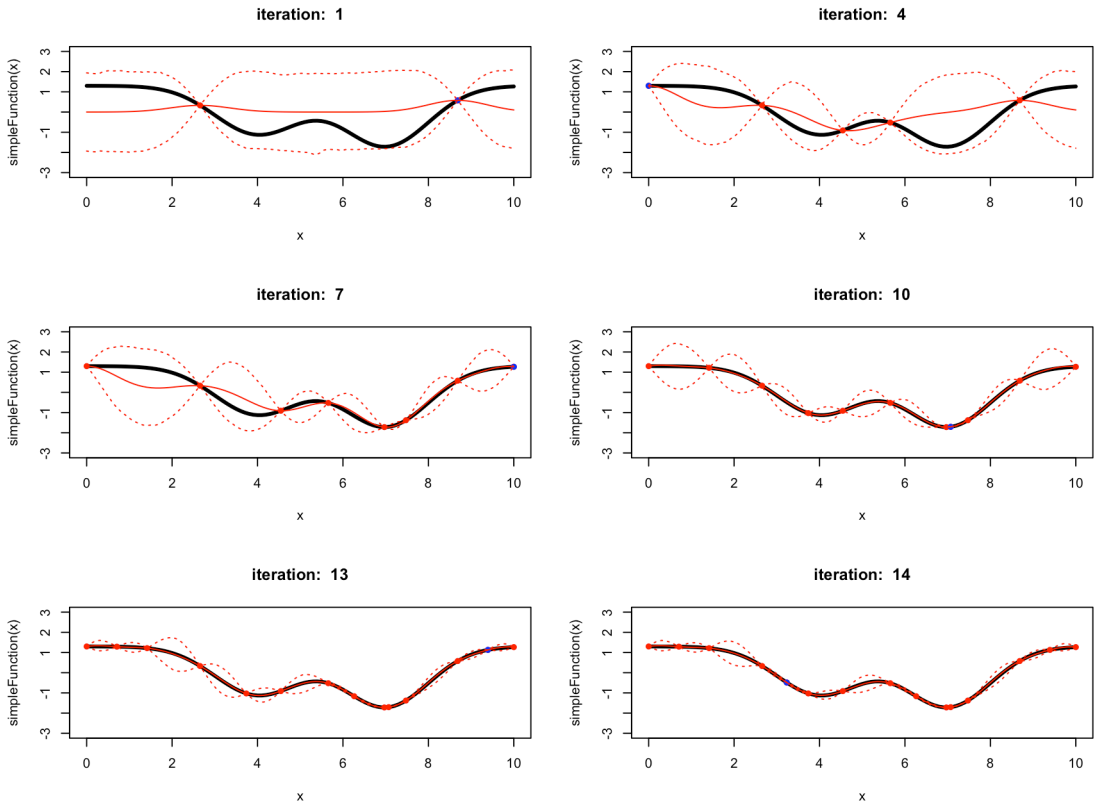

Figure A3: The evaluations at 15 evaluated locations.