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 essential extension of BO, particularly in addressing constraints.

The primary objective of the first study is to introduce how Bayesian optimization works, including Gaussian Process (GP) 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 is 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. GP model is frequently used to provide a posterior distribution to describe potential values at candidate locations. The posterior distribution is updated after we evaluate the new point. The acquisition function measures the potential of optimality of a new point by evaluating the objective function based on the current posterior distribution. This function combines two evaluation processes: exploitation and exploration. The former evaluates the level of improvement (comparing the current optimal point to the estimated value of the new point), and the latter evaluates the potential of a new candidate point (evaluating the uncertainty of the value of the new point). EI is the most commonly used acquisition function due to its interpretability and analysis form.

While EI is calculated based on 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 KG assesses the improvement based on the mean function rather than observation of the points. KG provides a small performance benefit in the standard BO problem with noise-free evaluation and substantial performance improvements in the problems with noise.[3]

Further, constraint BO (CBO) was briefly introduced in this paper. While the standard BO assumes the function's domain is simple and easy to assess membership, CBO adds functional constraints to the objective function to define 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 domain conditional on the new potential point added yet evaluated. This idea is similar to KG as the acquisition function is calculated at the stage where a new point is supposed to be added to the design, but we have not evaluated it. The "integration" refers to the improvement integrated out over the whole feasible set; this step requires knowing the weight of each design point. 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 function is known but differentiable, as it does not require calculating the derivative. BO can be extended to more complex optimization problems, for example, multi-task and multi-objective BO, where we have more than one objective functions.[4]

Although GP is used in most work on BO, 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 covariance matrix, if we employ empirical Bayes. On the other hand, the optimization algorithm will be expensively time-consuming if we employ a fully Bayesian approach, as 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.

3 Application

As the first paper provided the basic pseudo-code of BO, we will define a function to optimize using BO and the R built-in function `optimize` 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 denotes the standard normal density function. We begin by optimizing the objective function using `optimize`. Subsequently, we apply BO with the EI acquisition function and compare the outcomes of both approaches.

In this setup, the GP model includes only an intercept term set to zero. The covariance function is defined as the exponential of the negative squared Euclidean distance, i.e., $\exp -(x - x')^2$ for all $x, x' \in \mathcal{X}$. We assume noise-free evaluations, meaning the objective function is observed without measurement error.

Figure A1 illustrates the objective function (solid black curve), a set of sample functions (dashed grey curves) drawn from the prior Gaussian process before any observations are made, and the corresponding 95% credible intervals (dashed red lines). Notably, the objective function remains within the credible intervals across the entire domain, indicating that the surrogate model captures the potential range of the objective function well. This supports achieving the objective function values using the chosen surrogate model. The black horizontal and vertical line indicate the optimal point (local optimal) and its value obtained using the `R` built-in function.

As observations are collected, the uncertainty at the evaluated design points diminishes to zero, while it increases at points farther from the observations. With each additional evaluation, the surrogate model becomes increasingly aligned with the true objective function. This progression is illustrated in Figure A2. After 14 evaluations, the surrogate model closely approximates the objective function. The final optimization result is presented in Figure A3, showing that after 15 evaluations, the selected minimum corresponds to the global minimum of the objective function.

4 Discussion

Assuming noise-free evaluations, BO effectively identifies the global minimum. However, this assumption is often unrealistic, as real-world evaluations typically involve noise. In such cases, selecting the observed minimum to compute EI may yield inconsistent results. To address this, the second referenced paper recommends using the posterior mean instead of raw observations. While this study considers sequential (single-point) evaluations, the method can be extended to batch selection, which is more suitable when noise is present.

In CBO with the IECI acquisition function, the choice of the optimal point at the current stage must satisfy the monotonicity condition: $f_{\min} \leq \mathbb{E}[f(y \mid x)]$. See the second paper for details. Further, the selection of the covariance function should reflect prior beliefs about the function's smoothness and vary by application. A constant mean function (intercept only) is commonly used, allowing the data to inform the model structure; this is advantageous when prior knowledge about the model's trend is limited. However, challenges remain, including prior specification in fully Bayesian approaches and flat likelihood issues in empirical Bayes.

(`R` code of our application is described in detail in README file. Access this GitHub link for detail.)

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.

[4] Garnett R. *Bayesian Optimization*. Cambridge University Press; 2023.

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.

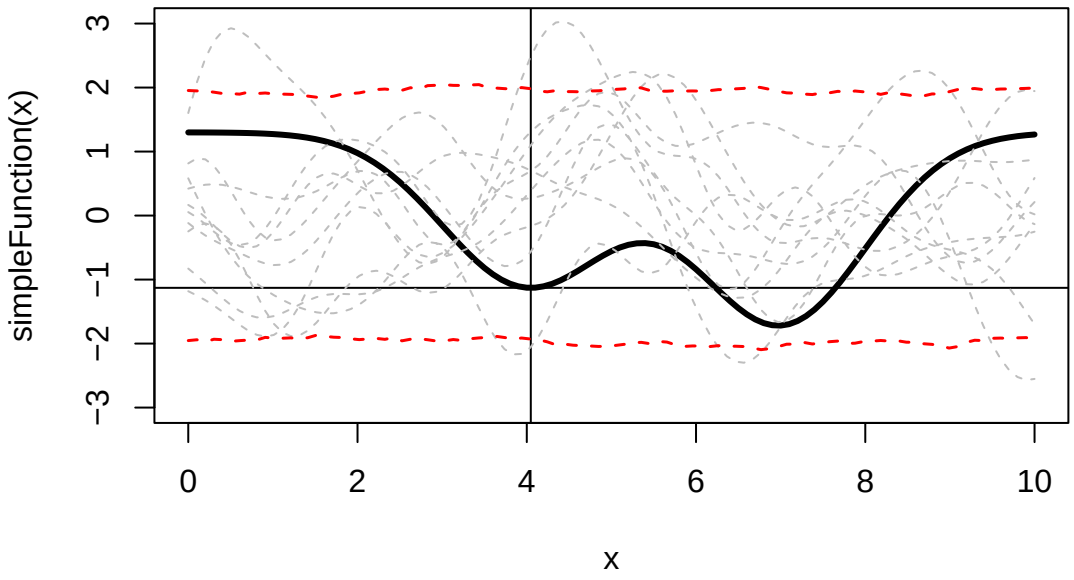

Figure A2: the plot shows the reduction of variance at evaluated locations.

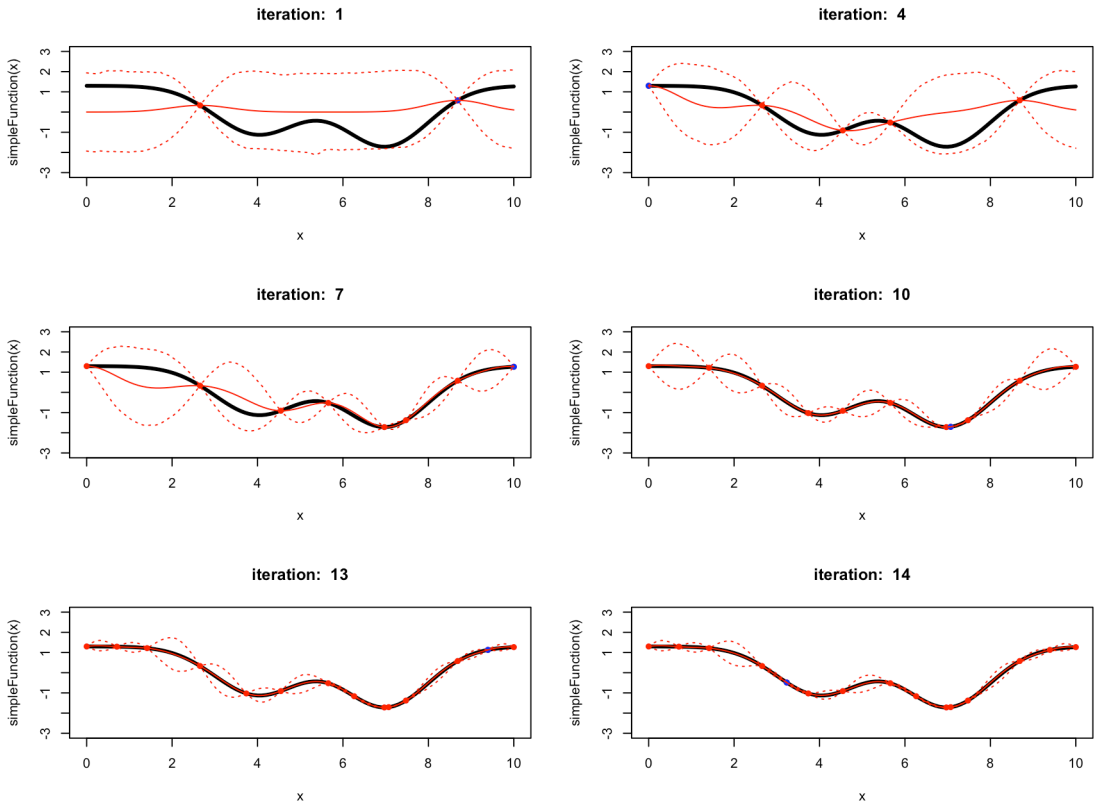

Figure A3: The evaluations at 15 evaluated locations.