Elementary Bayesian Optimization

응용수학특수연구III 기말 발표, 김선중

December 26, 2020

Table of Contents

1. References
2. Brief History of Bayesian Optimzation
3. An Overview of Bayesian Optimizaiton
4. A Few Elements of Bayesian Optimization
    1. Priors
    2. Covariance Function
    3. Acquisition function

1. References
2. Brief History of Bayesian Optimzation[3, 5]
3. An Overview of Bayesian Optimizaiton[3, 5]
4. A Few Elements of Bayesian Optimization[5]

References

1. James Bergestra, et al., 2011, “Algorithms for Hyper-Parameter Optimzation”
2. Jasper Snoek, et al., 2012, “Practical Bayesian Optimization of Machine Learning Algorithms”

References : Keywords in [1] and [2]

- Gaussian Process(GP)
- Expected Improvement(EI)
- Sequential Model Based Global Optimization(SMBO)
- Deep Belief Network (DBN)
- Tree-structured Parzen Estimator Approach(TPE)

References

3. Peter I. Frazier, 2018, “A Tutorial on Bayesian Optimization”
4. Peter I. Frazier, “TutORial : Bayesian Optimization”, `www.youtube.com/watch?v=c4KKvyWW_Xk&t=1924s`
5. Eric Brochu, et al., 2010, TR-2009 “A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning”

1. References
2. Brief History of Bayesian Optimzation[3, 5]
3. An Overview of Bayesian Optimizaiton[3, 5]
4. A Few Elements of Bayesian Optimization[5]

Brief History of Bayesian Optimzation

(Section 1 in [3], Section 2.5 in [5])

Bayesian optimization is originated with the work of Kushner, Žilinskas and Mockus.

- (Kushner, 1964) extended the method of sampling in finding maximum of a function. He extended to a new method of considering multiple sampling, which is proved to be useful than the former ones.
- In the paper (Žilinskas, 1975), the one-stage Bayesian method for one-variable minimzation (OBM) is suggested.

Brief History of Bayesian Optimzation

- (Močkus, 1978) together with Žilinskas et al, applied Bayesian approach to the problem of finding global maximum of multiextremal functions.
- (Močkus, 1975) also investigated Bayesian optimzation of multiextremal functions.
- These methodology is presented in his book “Bayesian Approach to Global Optimization” (Močkus, 1989).

Brief History of Bayesian Optimzation

Bayesian optimization received substantially more attention and popularized by Jones et al.

- In (Jones, 1998), he adopted response surface which is often called surrogate function, to reduce time and cost to approximate the given function. The mechanism is called efficient global optimization(EGO) algorithm.
- EGO is applied to derivative-free optimization and experimental design[5].

Brief History of Bayesian Optimzation

Following (Jones, 1998), innovations developed in that same literature includes

- multi-fidelity optimization (Huang, 2006; Sóbester, 2004)
- multi-objective optimization (Keane, 2006; Knowles, 2006; Močkus; Močkus, 1991)
- study of convergence rates(Calvin,1997; Calvin and Žilinskas, 2000; Calvin and Žilinskas, 2005; Calvin and Žilinskas, 1999)

Brief History of Bayesian Optimzation

The observation made by (Snoek, 2012) that BayesOpt is useful for training deep neural networks sparked a surge of interest within machine learning.

- multi-task optimization (Swersky, 2013; Toscano-Palmerin and Frazier, 2018)
- multi-fidelity optimization for DNN (Klein et al., 2016),
- parallel methods (Ginsbourger, 2007, 2010; Wang, 2016a; Wu and Frazier, 2016)
- Gaussian process regression (Kleinjnel, 2008 ; Salemi, 2014 ; Mehdad and Kleijnen, 2018)

1. References
2. Brief History of Bayesian Optimzation[3, 5]
3. An Overview of Bayesian Optimizaiton[3, 5]
4. A Few Elements of Bayesian Optimization[5]

An Overview of Bayesian Optimizaiton

- Bayesian optimization is a method for maximizing (or minimizing) a function $f : A \to \mathbb{R}$ where $A \subset \mathbb{R}^d$. Sometimes we write
$$\max_{x \in A} f(x).$$
- The objective function f is unknown, we are to sample from the domain A using surrogate function and acquisition function to achieve our goal.

An Overview of Bayesian Optimizaiton

In our problem, we are assuming the followings;

- d, the dimension of the domain, is assumed to be less than 20.
- x, the independent variable, is assumed to be within a simple domain A like k-cube $\prod_{i=1}^{k}[a_i, b_i]$.
- f is assumed to be a good function in a sense that it is (Lipschitz) continuous. But it is not differentiable. So it is impossible to approximate f using its derivatives or its second derivatives.
- Still, f is really complicated function, so one can not evaluate the value in low cost. And it lacks special structure like concavity or linearity.

Overview of Bayesian Optimization

- It is called *Bayesian* because it uses the famous "Bayes theorem"
$$P(M|E) \propto P(E|M)P(M).$$
- The posterior probability of a model M given evidence E is proportional to the likelihood of E given M, multiplied by the prior probability of M.
- Let x_i be the ith sample. Consider the set
$$D_{1:t} = \{(x_i, f(x_i))\,|1 \le i \le t\}$$
of ordered pair $(x_i, f(x_i))$, which plays a role of evidence. We have
$$P(f|D_{1:t}) \propto P(D_{1:t}|f)P(f).$$

Overview of Bayesian Optimization

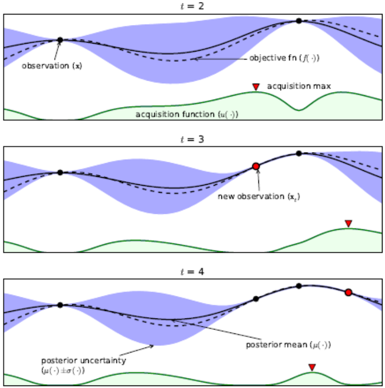


Figure 1: *An example of using Bayesian optimization on a toy 1D design problem. The figures show a Gaussian process (GP) approximation of the objective function over four iterations of sampled values of the objective function. The figure also shows the acquisition function in the lower shaded plots. The acquisition is high where the GP predicts a high objective (exploitation) and where the prediction uncertainty is high (exploration)—areas with both attributes are sampled first. Note that the area on the far left remains unsampled, as while it has high uncertainty, it is (correctly) predicted to offer little improvement over the highest observation.*

1. References
2. Brief History of Bayesian Optimzation[3, 5]
3. An Overview of Bayesian Optimizaiton[3, 5]
4. A Few Elements of Bayesian Optimization[5]

Priors

- We are to use Gaussian(normal) distribution for the prior, where we introduce the notion of *Gaussian process*.
- The collection $\{X_t : t \geq 0\}$ of random variables X_t where t is indexed in a continuous domain is called the random process.
- Gaussian process is a kind of random process, where each of the X_t for $t \geq 0$ is distributed normally.

Priors

- We are to approxiamte the function f.
- For each $x \in A$, the value $f(x)$ is not determined. So we may think of $f(x)$ as a random variable. That is, we may think of $\{f(x)|x \in A\}$ as a collection of random variable in a continuous domain A.
- Assuming $f(x)$ is distributed normally, we need only to specify its mean $m(x)$ for each $x \in A$ and covariance $k(x, x')$ for each $x, x' \in A$. We write

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')).$$

Priors

- Still, the function m and k is not specified yet.
- For convenience, let m be a zero function;

$$m(x) = 0. \quad (x \in A)$$

- We may choose k quite arbitrarily. But the most popular choice is to use squared exponential function;

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}||x_i - x_j||^2\right).$$

- Note that the value $k(x_i, x_j)$ converges to 1 as $(x_i - x_j)$ approaches to 0.

Covariance Function

- The (popular) kernel function we have defined is quite a naive one. We add a hyperparameter θ so that

$$k(x_i, x_j) = \exp\left(-\frac{1}{2\theta}||x_i - x_j||^2\right).$$

- For anisotropic models, we may set

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}(x_i - x_j)^T \mathrm{diag}(\theta)^{-2}(x_i - x_j)\right),$$

where diag stands for a diagonal matrix of a vector θ.

Covariance Function

- Another important kernel for Bayesian Optimization is the Matérn kernel (Matérn, 1960 ; Stein, 1999)
- It incorporates a smoothness parameter ζ to permit great flexibility in modelling functions:

$$k(x_i, x_j) = \frac{1}{2^{\zeta-1}\Gamma(\zeta)} \left(2\sqrt{\zeta}||x_i - x_j||\right)^{\zeta} H_{\zeta}\left(2\sqrt{\zeta}||x_i - x_j||\right),$$

where $\Gamma(\cdot)$ is the gamma function and $H_{\zeta}(\cdot)$ is the Bessel function of order ζ.

Acquisition Function

- The acquisition function plays a role of guiding the search for the optimum.
- It is defined so that high acquisition is related to *potentially* high values of the objective function f.
- Denote u by the acquisition function.
- At timestep t, we need to sample another point x_{t+1} from the domain. We choose x_{t+1} as

$$x_{t+1} = \text{argmax}_{x \in A} u(x|D).$$

- There are several kinds of acquisition functions, which we'll illustrate without deep explanation.

Acquisition Function : PI

- (Kushner, 1964) suggested maximizing *probability of improvement* over the incumbent $f(x^+)$ where

$$x^+ = \text{argmax}_{1\leq i\leq t} f(x_i),$$

so that

$$\begin{aligned} \text{PI}(x) &= P(f(x) \geq f(x^+)) \\ &= \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right). \end{aligned}$$

Here, Φ stands for the normal cumulative distribution function.

- Kushner's approach is a natural, but greedy one. It pursuits exploitation without exploration.

Acquisition Function : EI

- More satisfing alternative acquisition function is the one which considers not only the probability of improvement, but also the magnitude of the improvement a point can potentially yield.
- (Močkus, 1978) Let I be a function (called *improvement function*) defined by

$$I(x) = \max\{0, f_{t+1}(x) - f(x^+)\}.$$

Define the new point x at which we sample, by

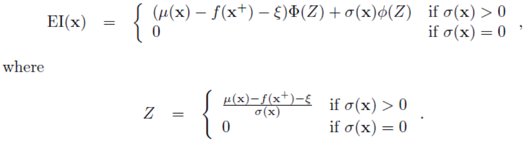
$$x = \text{argmax}_x \mathbb{E}\left(\max\{0, f_{t+1}(x) - f(x^+)\}|D_t\right).$$

Acquisition Function : Generalized EI

- EI, explained so far, can be evaluated analytically(Močkus, 1978 ; Jones, 1998) as follows;

$$
\begin{aligned}
\mathrm{EI}(\mathbf{x}) &= \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+))\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \\
Z &= \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma(\mathbf{x})}
\end{aligned}
$$

- (Lizotte, 2008) suggests new acquisition function that enables us to trade off between exploration and exploitation He adopted a parameter $\xi \geq 0$ such that



$$
\mathrm{EI}(\mathbf{x}) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases},
$$

where

$$
Z = \begin{cases} \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}.
$$

Thank you