The general scheme of Bayesian Optimization

Caglar Demir

Paderborn University

Abstract The aim of this paper is conveying the essence of Bayesian optimization. For reaching this purpose, the prerequisites of Bayesian optimization is thoroughly explained. Bayesian optimization utilizes the Bayes' theorem of setting a prior over the objective function and combining it with likelihood of evidence to get a posterior function. Therefore, the posterior function is used to construct an acquisition function in order to find extrema of objective function. At last, strengths and weaknesses of Bayesian optimization are explained.

Keywords: Bayes' Theorem, Gaussian distribution, Gaussian processes, Bayesian optimization

1 Introduction

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed as Arthur Samuel said once. Objective function and data are crucial part of learning process in Machine Learning. Roughly speaking, looking at data carefully and trying to understand the shape of objective function are the steps to follow in order to learn. Then, if it is the case, Machine Learning is dead end, when one does not have a closed-form expression for the objective function and evaluations are costly. Indeed, for vast the majority of the traditional machine learning algorithms fail in the case of absence of closed-form expression for objective function and requirement of many evaluations. If evaluations are costly and closed form expression for the objective function is not available, there are only few powerful techniques which work well for finding extrema of objective function and Bayesian optimization is one of them [2].

Bayesian Optimization (BO) is powerful strategy not only for finding the extrema of objective functions that are expensive to evaluate but for the problems whose objective function is non-convex [2] [6]. For many optimization problems, it is not an issue to find min of the objective function [4]. However, it is always the case that many evaluations on the objective function are required for finding extrema. In this regard, Bayesian optimization is some of the most efficient approaches [6]. It has been applying in many different fields. A popular application of BO is hyperparameter tuning [7], where the task is to minimize the validation error of a machine learning algorithm as a function of its hyperparameters. In Active learning , Bayesian optimization is also commonly used [2]. Bayesian Optimization is not only used in Computer Science but also in Petroleum industry in order to find the place to be drilled.

In this work, first contribution is conveying The general scheme of Bayesian Optimization with the help of examples for those who has not any prior knowledge about BO. The reader can expect undoubtedly that after reading this paper, one will derive solid background of Bayesian Optimization. To allow the essence of Bayesian Optimization to be seen, all prerequisites are explained. To achieve this purpose, Bayesian Theorem is explained and corresponding examples are given. It is indicated that how vital is to learn Bayesian Thinking, not only for computer scientists but for those who aim to learn logical thinking. Then, Gaussians are briefly explained since notion of Gaussians can not be comprehensively investigated in small amount of time. Gaussian Process is important technique in order to make a decision under uncertainty. BO assumes that the objective function was sampled from a Gaussian process and maintains a posterior distribution for this function as observations are made [7] [5]. Thus, Gaussian Process is used to select a prior over functions which express assumptions about objective function. Next, Bayesian Optimization is explained by and large. To reveal the essence of Bayesian Optimization, Multi-armed bandit problem is investigated by Bayesian perspective . As a conclusion, pros and cons of BO are concisely given.

2 What is Bayesian

Bayesian refers to methods which are related to statistical inference, named after Thomas Bayes. Bayesian view is fundamentally different than classical (frequentist) interpretation of probability. In Bayesian interpretation, probabilities of an event give a quantification of uncertainty [1]. Distinction between Bayesian interpretation and classical interpretation is drawn in next section with examples.

2.1 Bayes' theorem

Bayes' theorem describes the probability of an event, based on conditions that might be related to the other event. It can be seen as a way of understanding how the probability of a theory is affected by a new piece of incorporate evidence. Namely, the incorporate evidence converts the prior probability into the posterior probability.

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)} \tag{1}$$

In this paper, $p(.)$ indicates probability of corresponding event. The rule of probability states that, the joint distribution of two events $p(hd)$ is just equal to conditional $p(h \mid d)$ times marginal $p(d)$. Probability of h given d states that probability of h happening which is dependent on d. To be more clear, there is no ambiguity about event d, hence there is no probability of d. Fundamentally, the uncertainty in h can be evaluated after observing d in the form of posterior probability $p(h \mid d)$.

$p(h)$ is the prior which is an initial belief over an event. As an example, a child has a prior belief what a sheep is or how a sheep look like. The likelihood $p(d \mid h)$ is the quantity of incorporate evidences. In other words, it indicates new labeling, which is consistent with prior belief of sheep. Namely, it means that how many times our prior belief enhanced with an incorporate evidence. Combining the likelihood with the prior gives us the posterior $p(h \mid d)$. The posterior is basically final belief about an event. As a sheep example, after seeing more sheep, a child is more certain what a sheep is or how sheep looks like. Inherently, in the case of a child does not have an opinion about sheep, prior equals null and Bayesian interpretation is transformed into classical interpretation of probability. This transformation reveal that Bayesian Framework subsume classical interpretation, namely maximum likelihood. In the case of existing prior, prior belief about an event times likelihood of this event is proportional $\propto$ to posterior.

$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior} \tag{2}$$

Efficiency of Bayes' theorem stems from prior belief. Suppose, that a fair coin is tossed three times and lands heads each time. A classical interpretation would give 1 to the probability of landing heads which means that all tosses will heads. However, the Bayesian interpretation would give much less extreme probability with any reasonable prior [1]. The construction of prior is crucial in Bayes' theorem. In order to construct prior, Gaussian Process is widely used. To convey the essence of Gaussian Process, Gaussian distribution and Gaussian Process are explained in the follow sections.

3 Gaussian distribution

The theory of Gaussian distribution (GD) states that average of random variables which are independently drawn from independent distribution, converge in distribution, which becomes Gaussian distributed when the number of random variables is sufficiently large. A random variable is a variable whose possible values are numerical outcomes of a random phenomenon. GD is essentially a way of measuring the uncertainty for a variable which is continuous between $-\infty$ and ∞ . Formally, $x \in \mathbb{R}$ is, the univariate GD [1] is defined by

$$x \sim N(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2} - (x-\mu)^2\right) \tag{3}$$

the theorem states that x is a candidate from GD whose mean μ and variance σ^2 (variance is squared root of standard deviation). As one knows, the average value of a function $f(x)$ under probability distribution $p(x)$ is called expectation of a function $\mathbb{E}[x]$. Consequently, expectation of a function under the GD can be found as $\mathbb{E}[x] = \mu$ [1]. Maximum likelihood is commonly used for determining μ

and σ^2 by using the observed data set (To find more about maximizing likelihood function and maximize the log of the likelihood function [1]). In other words, to find the average temperature of Paderborn in July, one should have enough data and maximize the likelihood function with respect to μ.

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{4}$$

After calculation sample mean from the observed values, one can derive sample variance with respect to the sample mean.

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \tag{5}$$

As one can see that having enough data and deriving μ_{ML} and σ^2_{ML} is only requirement to make a proposition about GD. Moreover, sampling from GD is also possible. Knowing mean and variance enables one to state whether a data is sampled from corresponding GD which is explained later on this section.

A vector $X \in \mathbb{R}^2$, bivariate Gaussian distributed is defined by

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \tag{6}$$

X is simulated $\sim$ means that X is a candidate vector from GD with mean μ and covariance matrix Σ. In multivariate GD, variance is replaced with covariance matrix. Covariance matrix (equivalently correlation) shows the relation between points. Covariance matrix is a measure of how much two random variables change together. Correlation is any statistical relationship between two random variables which is shown by $corr(x_1, x_2) = \frac{cov(x_1,x_2)}{\sigma(x_1)\sigma(x_2)}$. The difference between covariance and correlation is that correlation is divided by standard deviations. More Precisely, $\Sigma_{12} = \mathbb{E}(x_1 x_2)$.

First plot which is the left hand side in figure 1 shows the probability density of x_1 and x_2. Namely, $P(x_1 x_2)$ pdf jointly models x_1 and x_2. In contrast the first plot, second plot exhibits the probability of x_2 happening is dependent on the value of x_1. To help visualization, whole contour plot[1] is cut where $x_1 = 1$.

An important property of the multivariate GD is shown in 1. Those plots indicate that *if two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian* [1]. Consequently, one can derive μ_{12} and Σ_{12} from jointly GD.

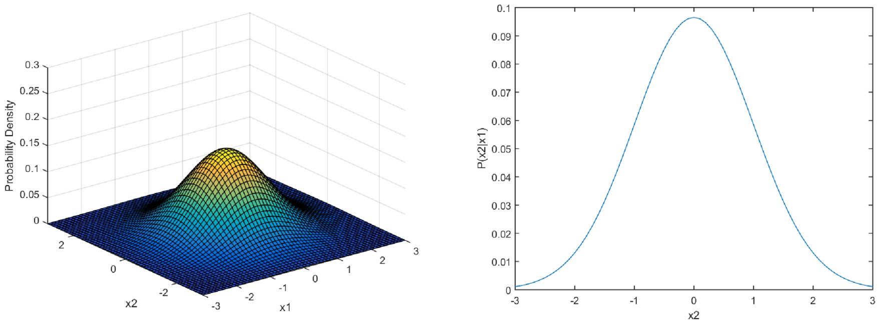
$$\mu_{12} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{7}$$

$$\Sigma_{12} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{8}$$

Fundamentally, the property of the bivariate GD allows us to derive conditional distribution from joint distribution. In other words, theorem allow us to go from big picture to a piece of big picture, which is called deductive process.

[1] Implementation of multivarient of Gaussian distribution https://goo.gl/vVt6Kj

Figure 1. Gaussian Distribution



Theorem also allows us to do reverse engineering. Namely, it is possible to make propositions from conditional distribution to joint distribution. The important aspect of this inductive process is to construct covariance matrix. In this work, covariance matrix is produced by squared exponential kernel. Besides, it should be indicated that there are many methods of measuring the similarities. However, squared exponential kernel is straightforward.

Squared exponential kernel $K_{ij} = e^{-|x_i - x_j|^2}$ squares the distance. In the case of $x_i = xj$ similarity equals one and in the case of $|x_i - xj| \to \infty$ similarity equals zero. After having training set and constructing covariance matrix by squared exponential kernel, any new point from test set can be predicted with mean and covariance. It means that from now one predict new points with high or low probability which enhances the one's certainty about a prediction.

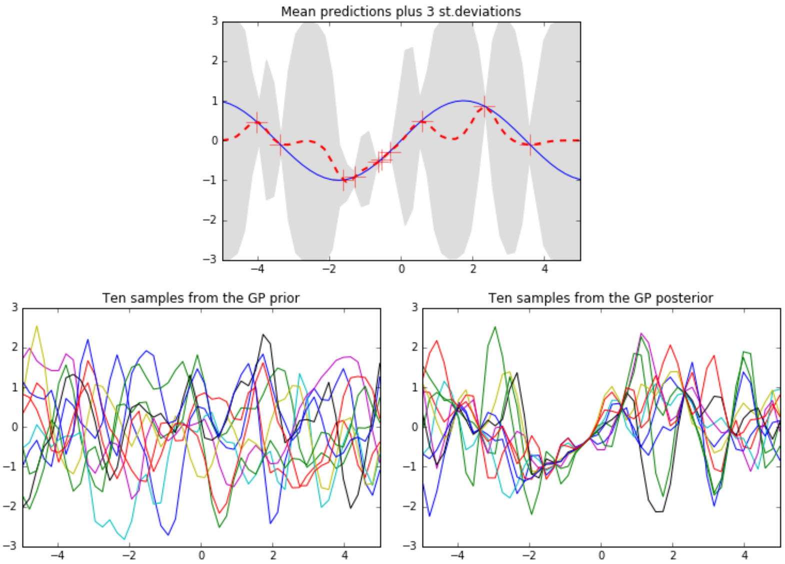
In this point, it should be highlighted that having the right expression for mean and variance, any new data can be predicted with its mean and variance. Therefore, if large number of data is available, a line can be plotted which crosses mean point of all data. In fact that, infinite number of points/functions can be used to draw a line. This is the idea of Gaussian Process.

3.1 Gaussian Process

A Gaussian process (GP) is a Gaussian distribution over functions [5]. Namely, GP is defined as a probability distribution over functions such that the set of values of a function evaluated at an arbitrary set of points jointly have a GD. It can be regared as an uncountable collection of random variables which have a joint Gaussian distribution [4]. The formula states that $f(x)$ is a candidate function from GP whose mean function $\mu(x)$ and covariance kernel $k(x, x\prime)$. The mean of function $\mu(x) = \mathbb{E}[f(x)]$ as one expect. The kernel function measures the similarities between functions which is shown as $k(x, x\prime) = \mathbb{E}[f(x)f(x\prime)]$ and equally $k(x, x\prime) = \exp(-\frac{1}{2l^2}(x - x')^2)$.

An important property of GP is that the joint distribution over N variables is specified by the mean and the covariance.

Figure 2. Gaussian Process [3]



Mean predictions plus 3 st.deviations

Ten samples from the GP prior

Ten samples from the GP posterior

In plot of Mean predictions, the blue line indicates the true unknown function which we aim to approximate. Red plus sign indicates training examples and red minus stands for mean prediction of test data which is noiseless. Therefore all training points are on the true unknown function. As theorem states, if two points close to each other, they are similar to each other with respect to kernel parameter which is used for measuring similarities and is defined by practitioner. Where the data accumulates, confidence intervals (grey areas) are squashed which is drawn by variance of functions. It states that the predictions can be in those areas with high certainty.

3.2 Effect of kernel hyperparameter

The value of kernel parameter has a crucial effect on Gaussian Process. In [2], it is indicated that it is necessary to add hyperparameter. The width of the kernel is controlled by hyperparameter l. Therefore Covariance function contains single hyperparameter $k(x, x') = \exp(-\frac{1}{2l^2}(x - x')^2)$. As it is explained before that the important part of the Kernel is the l which is called kernel parameter. Value of kernel parameter indicates what is area in which the points are regarded as similar points. Choosing the kernel parameter carries huge importance on the success of Gaussian.

Figure 3. Kernel Parameters 0.1 and 1

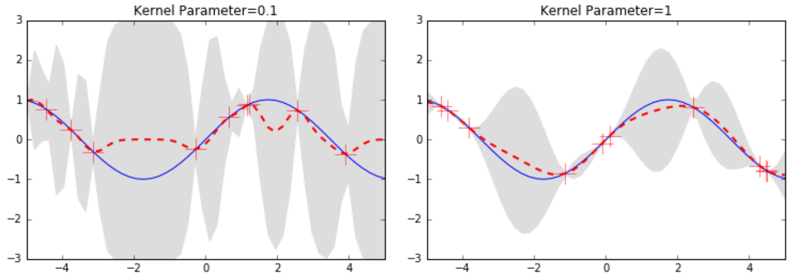
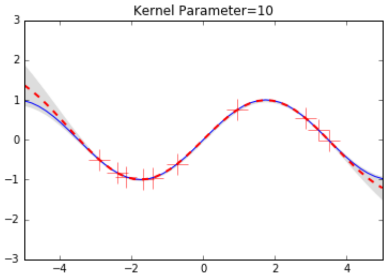


Figure 4. Kernel Parameter =10



3.3 Bayesian Framework and Gaussian

After explaining Bayesian Framework and Gaussian separately, it would be helpful to give a quick summary about all process. First assumption is that data is Gaussian distributed. This is the assumption which has to be made in order to use Gaussian. As an brief example, it is true to state that probability of a person whose height 180 cm. and whose weight is 75 kg, higher than probability of a person whose height 220 cm and weight is 75 kg. Height and weight are random variables and their distribution converge Gaussian distributed with sufficient number of them. Using Gaussian enables to define likelihood, specify prior and computer the posterior for Bayesian. Same process is happening also in GP with only difference. Instead of using vectors, functions are using, namely infinite number of points. Gaussian is preferable to use as a statistical model due to flexibility and tractability. In Bayesian optimization, only two additional concepts are explained which are acquisition function and extrema of objective function. In Bayesian optimization, it is aimed not to figuring out the whole objective function but only extrema of it. To achieve this goal, acquisition function is utilized.

4 Bayesian Optimization

Bayesian optimization (BO) is a form of optimization, when one does not have a closed-form expression for the objective function but where one can obtain observations of this function at sampled values [2] [6]. In our paper, Bayesian Optimization is used to find max of objective function. Bayesian optimization typically works by assuming the unknown function was sampled from a Gaussian process and maintains a posterior distribution for this function as observations are made [7]. BO constructs a probabilistic model (GP) for $f(x)$ and at each iteration the model is used to select the most promising candidate for evaluation [4]. This feature of BO enables us to find the extrema of non-convex functions with relatively few evaluations. Efficiency of BO comes from the incorporate prior belief regarding the direct sampling and exploration-exploitation trade off. It is called Bayesian due to utilizing Bayes' Theorem which is explained in section 1.

Algorithm 1 Bayesian Optimization

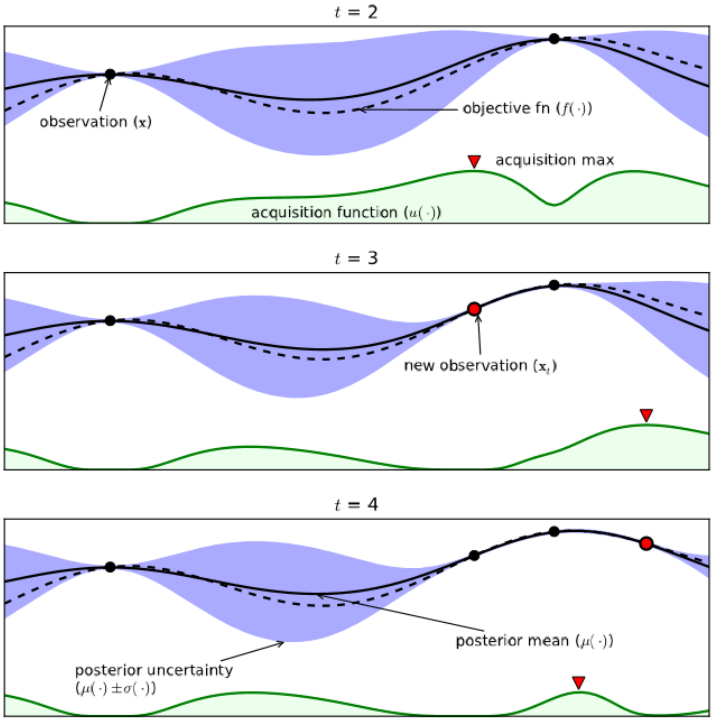
```
for t=1,2,.... do
    Find $x_t$ by posterior and $x_t = argmax_x u(x|D_{1:t-1})$
    Sample the objective function:$y_t = f(x_t) + \epsilon_t$.
    Augment the data $D_{1:t} = D_{1:t-1}, (x_t, y_t)$
and update the GP
end for
```

There are two major choices that must be made when one utilizes Bayesian optimization. First, one must select a prior over functions that will express assumptions. For this GP is chosen, due to its flexibility [5] and tractability [2]. Second, one must choose the acquisition function, which is used to construct a utility function from the posterior.

The acquisition function allows us to determine the next point to evaluate [2]. It is important to notice that after having two observation, exploration starts around the observation which is highest observation on the function.

Since the aim is to find the max of the objective function, exploration starts

Figure 5. Bayesian Optimization with multiple points [2]



around of the point which is right hand side in figure 5. After finding highest observation, second exploration-exploitation criteria comes in which aims to gain more information with every sampling.
In other words, where the uncertainty is high (the max of acquisition function), exploring those areas inherently gives higher information. After every sampling, Gaussian is refitted in order to update current knowledge and reshape the posterior. As one might wonder whether finding the max of acquisition function is also new optimization problem. Indeed, BO carries a new optimization problem with itself.

4.1 Bayesian optimization and Multi-armed bandits

The Multi-armed bandit machine problem is a problem in which a gambler at a row of slot machines has to decide which machines to play, how many times to play each machine. Moreover, it is assumed that one has some trials to play which is confined by either time or money. One is not allowed to play forever. If it is the case that there is only one bandit machine to play with in order to gain reward (money), there is no exploration exploitation trade-off. One can not explore due to no available extra bandit machine. However if there are some bandit machine, exploration-exploitation dilemma occurs and notion of regret comes in. Essentially, notion of regret is equal to reward of player minus reward of best action which is the case most of the time. In the case of winning on the machine is satisfying as long as no one wins bigger reward. If it is the case that one gets bigger reward, exploration of other bandit machines is getting more attractive. From the perspective BO, keep playing on the same machine has the highest priority as long as getting higher reward than everyone. However if it is the case that someone gets bigger reward, then the next step to take is dependent upon the settings of Acquisition Functions. One either stays and keeps playing on the same machine or starts exploring other bandit machines.

4.2 Acquisition Function and Exploration-Exploitation

Exploration-exploitation trade-off is balanced with Acquisition function is which is $\mu(x) + k\sigma(x)$ [2]. Exploration-exploitation trade-off is essentially saying that either keep playing on the same machine (in the case of winning, otherwise surely exploitation is the only choice) whilst none gets bigger reward or keep exploiting in order to find bigger reward than others. It should be pointed out again that Regret is equal to one's reward minus reward of best action. As long as getting higher reward than others, keep playing is the first choice.

Finding next point to be chosen or next machine to be played on where the mean (exploitation) and variance (exploration) are high, is new optimization problem. BO carries an extra optimization problem with itself. However finding next point is relatively easy problem to be optimized. Result of solving this problem varies according to setting of practitioner. If practitioner regards exploring more important, then constant k is relatively large. It is relatively large due to data and importance of exploring. On the other hand, If practitioner regards exploitation more important, then k should be small.

5 Conclusion

With this work, it is given the general scheme of Bayesian Optimization. Bayes' theorem are thoroughly explained. Gaussian Distribution and Gaussian Process are introduced and visualized. At least but not last, Bayesian Optimization is investigated. However, the disadvantages of using Bayesian Optimization is not mentioned. There are some drawbacks of using Bayesian Optimization. Firstly, Bayesian Optimization is only useful if evaluations are costly. If it is not the case, Bayesian Optimization is not preferable. Constructing prior is critical to solve optimization problem. Gaussian processes are not always the best or easiest solution, but even when they are,desing of kernel requires expertise [2]. Importance of kernel parameter is already seen. In the case of choosing wrong kernel parameter, Bayesian Optimization will not give satisfying result. Moreover, the trade-off between exploration and exploitation in the acquisition function depends too much on practitioners. It is often unclear how to handle the trade-off between exploration and exploitation in the acquisition function. Giving high priority to exploitation can lead local maximization and too much exploration leads less improvement of movement.

References

1. CM Bishop. Bishop pattern recognition and machine learning, 2001.
2. 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. *CoRR*, abs/1012.2599, 2010.
3. Nando de Freitas. Gaussian processes for nonlinear regression. 2013.
4. Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John Cunningham. Bayesian optimization with inequality constraints. In *ICML*, pages 937–945, 2014.
5. Kenji Kawaguchi, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Bayesian optimization with exponential convergence. In *Advances in Neural Information Processing Systems*, pages 2809–2817, 2015.
6. Jonas Mockus. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.
7. Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.