High Dimensional Bayesian Optimization with Elastic Gaussian Process - Project Report

Sai Niranjan
BS Mathematics
SR : 15543

I. Introduction

As part of my reading project for the course MA 363:Probability in High Dimensions, I chose the paper *High Dimensional Bayesian Optimization with Elastic Gaussian Process*,published in ICML,2017. Bayesian optimization is an efficient way to optimize expensive black-box functions such as designing a new product with highest quality or hyperparameter tuning of a machine learning algorithm. However, it has a serious limitation when the parameter space is high-dimensional as Bayesian optimization crucially depends on solving a global optimization of a surrogate utility function in the same sized dimensions. The surrogate utility function, known commonly as acquisition function is a continuous function but can be extremely sharp at high dimension - having only a few peaks marooned in a large terrain of almost flat surface. The authors propose an innovative means for the same which enables local gradient-dependent algorithms to move through the flat terrain by using a sequence of gross-to-finer Gaussian process priors on the objective function. Both theoretical guarantees and experimental verification are presented in the original work but due to space constraints only the theoretical guarantees are shown here.

II. Background

Following the original paper, we present a basic overview of the mathematical background involved.

A. *Gaussian Process:*

A Gaussian process (GP) is a distribution over functions specified by its mean $m(.)$ and covariance kernel function $k(.,.)$. Let $\mathbf{x_{1:t}}$ be a set of observations with a vector of response values $\mathbf{f(x_{1:t})}$ we have that;

$$\mathbf{f(x_{1:t})} \sim \mathcal{N}(\mathbf{m(x_{1:t})}, \mathbf{K(x_{1:t}, x_{1:t})}) \tag{1}$$

where,

$$\mathbf{K}(\mathbf{x}_{1:t}, \mathbf{x}_{1:t}) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_t) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_t, \mathbf{x}_1) & \cdots & k(\mathbf{x}_t, \mathbf{x}_t) \end{bmatrix}$$

k is a kernel function here. The choice of the kernel depends on prior beliefs about smoothness properties of the objective function. A popular kernel function is the squared exponential (SE) function, which is defined as :

$$k((\mathbf{x_i}, \mathbf{x_j})) = \mathbf{exp}(\frac{\mathbf{1}}{\mathbf{2l}} \|\mathbf{x_i} - \mathbf{x_j^2}\|) \tag{2}$$

l here is a parameter quantifying smoothness. The predictive distribution of a GP is tractable analytically. For a new point ,$\mathbf{x_{t+1}}$ the joint probability distribution of the known values $\mathbf{f_{1:t}} = \mathbf{f(x_{1:t})}$ and the predicted function value $\mathbf{f_{t+1}}$ is given by

$$\begin{pmatrix} \mathbf{f}_{1:t} \\ f_{t+1} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}(\mathbf{x}_{1:t}) \\ m(\mathbf{x}_{t+1}) \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{x},\mathbf{x}) & \mathbf{k} \\ \mathbf{k}^T & k(\mathbf{x}_{t+1}, \mathbf{x}_{t+1}) \end{bmatrix} \right)$$

where,$\mathbf{k} = [\mathbf{k}(\mathbf{x_{t+1}}, \mathbf{x_1}), ...\mathbf{k}(\mathbf{x_{t+1}}, \mathbf{x_t})]^{\mathbf{T}}$],also, $\mathbf{K}(\mathbf{x}, \mathbf{x}) = \mathbf{K}(\mathbf{x_{1:t}}, \mathbf{x_{1:t}})$, we can simplify the calculation by letting, $\mathbf{m}(\mathbf{x_{1:t}}) = \mathbf{0}$.The predictive distribution of f_{t+1} then can be represented by:

$$f_{t+1} \mid \mathbf{f}_{1:t} \sim \mathcal{N}(\mu_{t+1}(\mathbf{x}_{t+1} \mid \mathbf{x}_{1:t}, \mathbf{f}_{1:t}), \sigma^2_{t+1}(\mathbf{x}_{t+1} \mid \mathbf{x}_{1:t}, \mathbf{f}_{1:t}))$$

where, $\mu_{t+1}(.) = \mathbf{k^T K^{-1} f_{1:t}}$ and $\sigma^2_{t+1} = k(\mathbf{x_{t+1}}, \mathbf{x_{t+1}}) - \mathbf{k^T K^{-1} k}$.

B. *Bayesian Optimization:*

A traditional optimization problem is to find the maximum or minimum of a function f(x) over a compact domain X . In real applications such as hyperparameter tuning for a machine learning model or experiments involving making of physical products, f(x) is unknown in advance and expensive to evaluate. Bayesian optimization (BO) is a powerful tool to optimize such expensive black-box functions. A common method to model the unknown function is using a Gaussian process as a prior. The posterior is maintained based on observations and allows prediction for expected function values at unseen locations. An *acquisition function* $a(\mathbf{x}|\mathbf{x_{1:t}}, \mathbf{f_{1:t}})$ is used to steer the search towards an optimum. The acquisition functions considered for analysis in the paper are *EI* and *UCB* functions.

The EI-based acquisition function computes the expected improvement with respect to the current maximum $f((\mathbf{x}^+))$. We have the **Improvement function** defined to be,

$$\mathbf{I}(\mathbf{x}) = \max\{\mathbf{0}, \mathbf{f_{t+1}}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^+)\} \tag{3}$$

where,$f_{t+1}(\mathbf{x})$ has mean $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$ as per the GP. EI or the **Expected Improvement** is then computed as,

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

where $Z(x) = \frac{\mu(\mathbf{x}) - \mathbf{f}(\mathbf{x}^+)}{\sigma(x)}$ and $\Phi(Z)$ and $\phi(Z)$ are the pdf and cdf of the standard normal distributions, evaluated at Z. The *UCB* Acquistion function is defined to be,

$$UCB(\mathbf{x}) = \mu(\mathbf{x}) + \nu\sigma(\mathbf{x}) \tag{4}$$

, here ν represents an increasing sequence of positive numbers.

In each iteration of Bayesian optimization, the most promising $\mathbf{x_{t+1}}$ is found by maximizing the acquisition function and then y_{t+1} is evaluated. The new observation is augmented to update the GP which is in turn is used to construct a new acquisition function. These steps are repeated till a satisfactory outcome is reached or the iteration limit is achieved.

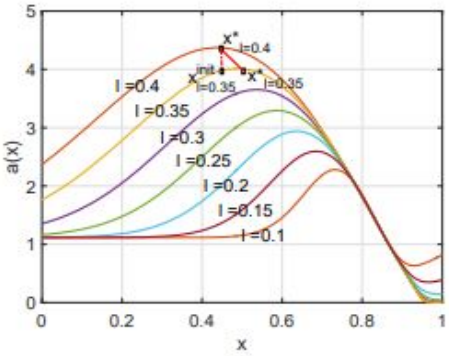
III. Motivation for using Elastic Gaussian Processes:

Current methods broadly involve using the *gradients* or going derivative free. Gradient methods are impractical in higher dimensions due to the geometry of the acquisition functions, many of them have large flat regions and random initialization into these regions makes the gradient based approaches unable to perform. On the other hand, derivative-free methods while safe from this issue, rely on heuristics and leverage properties such as Lipschitz continuity, the runtime however of such methods tend to become *exponential* versus the number of dimensions, thus going to higher dimensions is a big issue.

To model $f(\mathbf{x})$, the authors use a Gaussian process with zero mean as a prior and the SE kernel as the covariance function with a target length-scale l . The target length-scale can be set by the user or can be separately inferred by using the MLE method for example. We have that;

$$\begin{aligned} \mu(\mathbf{x}) &= \mathbf{k}^T\mathbf{K}^{-1}\mathbf{y} \\ \sigma^2(\mathbf{x}) &= 1 - \mathbf{k}^T\mathbf{K}^{-1}\mathbf{k} \end{aligned}$$

Clearly, the acquisition function depends on the GP kernel length-scale l. The importance of this dependency can be seen from the below figure;



As we can see changing the smoothness parameter l ends up presenting different terrains. As can be seen, when the length-scale is low, some portions of the parameter space are flat. This observation is very crucial in high dimensions. For example, the acquisition function with length-scale 0.1 is extremely flat. However when the length-scale is above 0.2, the acquisition functions starts to have significant gradients. Crucially, the authors show in their work that the optimal solutions for different l values are located very close to each other.

Following these observations, an elegant solution is presented by the authors. Using a large enough length-scale modify the terrain so that the derivative is no longer insignificant at any point in the domain. Now, using the optimum here as an initialization, slowly reduce the length scale until the original one is reached where the optimum can be found much more easily. Therefore, it remains to be shown that such a modification is valid, and further that the difference in the acquisition function is smooth with respect to the change in length-scale. This implies that the extrema of the consecutive acquisition functions are close but different only due to a small difference in the length-scales. This method denoted as *Elastic GP* by the authors has been presented formally in the following algorithms:

Algorithm 1 High Dimensional Bayesian Optimization with Elastic Gaussian Process

1: **for** $t = 1, 2 \cdots$ **do**
2: Sample the next point $\mathbf{x}_{t+1} \leftarrow \text{argmax}_{\mathbf{x}_{t+1} \in \mathcal{X}} a(\mathbf{x} \mid \mathcal{D}_{1:t}, l)$ using Alg. 2
3: Evaluate the value y_{t+1}
4: Augment the data $\mathcal{D}_{1:t+1} = \{\mathcal{D}_{1:t}, \{\mathbf{x}_{t+1}, y_{t+1}\}\}$
5: Update the kernel matrix $\mathbf{K}$
6: **end for**

Algorithm 2 Optimizing acquistion function using EGP

Input: a random start point $\mathbf{x}_{init} \in \chi$, the length-scale interval $\triangle l, l = l_\tau$.

1: Step 1:
2: **while** $l \leq l_{max}$ **do**
3: Sample $\mathbf{x}^* \leftarrow \text{argmax}_{\mathbf{x}^* \in \mathcal{X}} a(\mathbf{x} \mid \mathcal{D}_{1:t}, l)$ starting with $\mathbf{x}_{init}$;
4: **if** $||\mathbf{x}_{init} - \mathbf{x}^*|| = 0$ **then**
5: $l = l + \triangle l$
6: **else**
7: $\mathbf{x}_{init} = \mathbf{x}^*$, break;
8: **end if**
9: **end while**
10: Step 2:
11: **while** $l \geq l_\tau$ **do**
12: $l = l - \triangle l$
13: Sample $\mathbf{x}^* \leftarrow \text{argmax}_{\mathbf{x}^* \in \mathcal{X}} a(\mathbf{x} \mid \mathcal{D}_{1:t}, l)$ starting with $\mathbf{x}_{init}$;
14: **if** $||\mathbf{x}_{init} - \mathbf{x}^*|| = 0$ **then**
15: $\triangle l = \triangle l/2$
16: **else**
17: $\mathbf{x}_{init} = \mathbf{x}^*$
18: **end if**
19: **end while**

Output: the optimal point $\mathbf{x}_{t+1} = \mathbf{x}^*$ to be used in Alg.1

IV. Theoretical Analysis

We first want to show that gradient of the acquisition functions becomes significant beyond a certain l so that our algorithm can find an optimal solution compared to any start point. The authors then prove the following lemma:

Lemma 1. $\exists l : \left\| \frac{\partial a(\mathbf{x})}{\partial \mathbf{x}} \right\|_2 \geq \varepsilon$ *for* $l_\tau \leq l \leq l_{max}$.

Proof. The Lemma can be proved if we prove that $\left| \frac{\partial a(\mathbf{x})}{\partial x_i} \right| \geq \varepsilon, \forall i$. We consider both forms: UCB and EI.

For the UCB case we get;

the partial derivative of UCB can be written as

$$\begin{aligned} \frac{\partial a(\mathbf{x})}{\partial x_i} &= \frac{\partial \mu(\mathbf{x})}{\partial x_i} + \nu \frac{\partial \sigma(\mathbf{x})}{\partial x_i} \\ &= \frac{\partial \mathbf{k}^T}{\partial x_i} \mathbf{K}^{-1} \mathbf{y} + \frac{\nu}{\sigma(\mathbf{x})} (-\frac{\partial \mathbf{k}^T}{\partial x_i} \mathbf{K}^{-1} \mathbf{k}) \end{aligned}$$

The $\frac{\partial \mathbf{k}^T}{\partial x_i}$ is dependent on the form of the covariance function: it is $1 \times t$ matrix whose $(1, j)^{th}$ element is $\frac{\partial cov(\mathbf{x}, \mathbf{x}_j)}{\partial x_i}$. For the SE kernel

$$\begin{aligned} \frac{\partial cov(\mathbf{x}, \mathbf{x}_j)}{\partial x_i} &= \exp\left(-\frac{||\mathbf{x} - \mathbf{x}_j||^2}{2l^2}\right)\left(-\frac{(x_i - x_{ji})}{l^2}\right) \\ &= -\frac{d_{ji}}{l^2} cov(\mathbf{x}, \mathbf{x}_j) \end{aligned}$$

where $d_{ji} = x_i - x_{ji}$.

To simplify the proof, we assume that we have the worst case that only one observation $\mathbf{x}_0$ exists and thus

$$\begin{aligned} &\frac{\partial a(\mathbf{x})}{\partial x_i} \\ &= \frac{y_0 \partial cov(\mathbf{x}, \mathbf{x}_0)}{\partial x_i} - \frac{vcov(\mathbf{x}, \mathbf{x}_0)}{\sqrt{1 - cov^2(\mathbf{x}, \mathbf{x}_0)}} \frac{\partial cov(\mathbf{x}, \mathbf{x}_0)}{\partial x_i} \\ &= -\frac{d_{0i}}{l^2} cov(\mathbf{x}, \mathbf{x}_0) y_0 + \frac{vcov(\mathbf{x}, \mathbf{x}_0)}{\sqrt{1 - cov^2(\mathbf{x}, \mathbf{x}_0)}} \frac{d_{0i}}{l^2} cov(\mathbf{x}, \mathbf{x}_0) \\ &= \frac{d_{0i}}{l^2} \left(\frac{vcov^2(\mathbf{x}, \mathbf{x}_0)}{\sqrt{1 - cov^2(\mathbf{x}, \mathbf{x}_0)}} - cov(\mathbf{x}, \mathbf{x}_0) y_0 \right) \end{aligned}$$

$cov(\mathbf{x}, \mathbf{x_0})$ is very small as $\|\mathbf{x} - \mathbf{x_0}\| >> \mathbf{0}$. Therefore, we get:

$$\begin{aligned} \frac{\partial a(\mathbf{x})}{\partial x_i} &= -\frac{d_{0i}}{l^2} cov(\mathbf{x}, \mathbf{x}_0) y_0 \\ &= -\frac{d_{0i} y_0}{l^2} \exp\left(-\frac{||\mathbf{x} - \mathbf{x}_0||^2}{2l^2}\right) \end{aligned}$$

We rewrite it as

$$\left| \frac{\partial a(\mathbf{x})}{\partial x_i} \right| = \frac{\alpha_1}{l^2} \exp\left(-\frac{\alpha_2}{l^2}\right)$$

where $\alpha_1 = |d_{0i} y_0|$ and $\alpha_2 = ||\mathbf{x} - \mathbf{x}_0||^2/2$.

To have $\left| \frac{\partial a(\mathbf{x})}{\partial x_i} \right| \geq \varepsilon$, the equation $\exp\left(-\frac{\alpha_2}{l^2}\right) \geq \frac{\varepsilon l^2}{\alpha_1}$ must hold for a l between $l_\tau \leq l \leq l_{max}$. In fact, we can find a l to hold the inequality since $\exp\left(-\frac{\alpha_2}{l^2}\right)$ is a decreasing function with the range $(0, 1]$ whilst $\frac{\varepsilon l^2}{\alpha_1}$ is an increasing function with the range $(0, +\infty)$ by considering l_τ can approach to 0 and l_{max} can approach to infinity in theory. Therefore Lemma 1 has been proved for the UCB acquisition function.

Now, for the EI case we have;

the partial derivative can be written as

$$\frac{\partial a(\mathbf{x})}{\partial x_i} = [Z\Phi(Z) + \phi(Z)]\frac{\partial \sigma(\mathbf{x})}{\partial x_i} + \sigma(\mathbf{x})\Phi(Z)\frac{\partial Z}{\partial x_i}$$

where $Z = \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma(\mathbf{x})}$ and

$$\frac{\partial \sigma(\mathbf{x})}{\partial x_i} = -\left(\frac{\partial \mathbf{k}^T}{\partial x_i}\mathbf{K}^{-1}\mathbf{k}\right)/\sigma(\mathbf{x})$$

$$\frac{\partial Z}{\partial x_i} = \left(\frac{\partial \mathbf{k}^T}{\partial x_i}\mathbf{K}^{-1}\mathbf{y} - Z\frac{\partial \sigma(\mathbf{x})}{\partial x_i}\right)/\sigma(\mathbf{x})$$

therefore,

$$\frac{\partial a(\mathbf{x})}{\partial x_i} = -\phi(Z)\left(\frac{\partial \mathbf{k}^T}{\partial x_i}\mathbf{K}^{-1}\mathbf{k}\right)/\sigma(\mathbf{x}) + \Phi(Z)\frac{\partial \mathbf{k}^T}{\partial x_i}\mathbf{K}^{-1}\mathbf{y}$$

Substituting from our result on partial derivative of cov and taking similar assumptions as in UCB, we get,

$$\begin{aligned}
&\frac{\partial a(\mathbf{x})}{\partial x_i} \\
&= \frac{y_0\Phi(Z)\partial cov(\mathbf{x}, \mathbf{x}_0)}{\partial x_i} - \frac{\phi(Z)cov(\mathbf{x}, \mathbf{x}_0)}{\sqrt{1 - cov^2(\mathbf{x}, \mathbf{x}_0)}}\frac{\partial cov(\mathbf{x}, \mathbf{x}_0)}{\partial x_i} \\
&= \frac{d_{0i}}{l^2}\left(\frac{\phi(Z)cov^2(\mathbf{x}, \mathbf{x}_0)}{\sqrt{1 - cov^2(\mathbf{x}, \mathbf{x}_0)}} - cov(\mathbf{x}, \mathbf{x}_j)y_0\Phi(Z)\right)
\end{aligned}$$

As $\phi(Z)$ lies in $[0, 1]$ we can ignore the first term to get;

$$\frac{\partial a(\mathbf{x})}{\partial x_i} = -\frac{d_{0i}}{l^2}\exp\left(-\frac{||\mathbf{x} - \mathbf{x}_0||^2}{2l^2}\right)y_0\Phi\left(\frac{\mu(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma(\mathbf{x})}\right)$$

As l increasing, $\mu(\mathbf{x})$ becomes smaller and $\sigma(\mathbf{x})$ becomes larger and then $\Phi(\cdot) \to 1$. The equation above becomes similar with Eq.(3.3). Therefore Lemma 1 is proved for EI.

In the second step of our algorithm (seen in Step 2 of Alg.2), our purpose is to find $\triangle l$ which makes the start point of the local optimizer move to a finer region. We need to show that

$$\left|\frac{\partial a(\mathbf{x})}{\partial x_i}\bigg|_{l=l^*} - \frac{\partial a(\mathbf{x})}{\partial x_i}\bigg|_{l=l^*+\triangle l}\right| \leq \varepsilon, \ for \ \triangle l < \delta$$

It is directly related to $\frac{\partial a(\mathbf{x}|\mathcal{D}_{1:t}, l)}{\partial \mathbf{x}}$ being smooth. The following lemma guarantees that.

Lemma 2. *$g(\mathbf{x}, l)$ is a smooth function with respect to l, where $g(\mathbf{x}, l) = \frac{\partial a(\mathbf{x}|\mathcal{D}_{1:t}, l)}{\partial \mathbf{x}}$.*

For UCB we compute the derivative of $g(\mathbf{x}, \mathbf{l})$ with respect to l to get;

$$\begin{aligned}
\frac{\partial g(x_i, l)}{\partial l} &= \frac{2d_{0i}y_0}{l^3}\exp\left(-\frac{||\mathbf{x} - \mathbf{x}_0||^2}{2l^2}\right) \\
&+ \frac{d_{0i}y_0}{l^2}\exp\left(-\frac{||\mathbf{x} - \mathbf{x}_0||^2}{2l^2}\right)\frac{||\mathbf{x} - \mathbf{x}_0||^2}{l^3}
\end{aligned}$$

Apparently, $\frac{\partial g(x_i, l)}{\partial l}$ is continuous in the domain of l. Therefore, $g(\mathbf{x}, l)$ is a smooth function with respect to l. The similar proof can be done for EI.

V. Conclusion:

The authors then proceed to evaluate their method on three different benchmark test functions and two real-world applications which are training cascaded classifiers and for alloy composition optimization. The results are promising and empirically verify the claims that have been proved here earlier. Because of space constraints the results are not shown here, but the reader may very well chose to go through them at their leisure from the original paper. The main aim of this report was to discuss a novel algorithm that had been proposed for performing Bayesian Optimization in High Dimensional spaces. At high dimensions, as we discussed earlier, the acquisition function becomes very flat on a large region of the space causing gradient-dependent methods to fail.Following the authors, we prove that a) Gradients can be induced by increasing the length-scales of the GP prior and b) Acquisition functions which differ only due to small difference in length-scales have values close to each other. Based on these the authors formulate their algorithm that first finds a large enough length-scale to enable the gradient-dependent optimizer to perform, and then gradually reduces the length-scale while also sequentially using the optimum of the larger length-scale as the initialization for the smaller. The author of this study chose this topic due to the relevance of the topic to the course attended and further based on research work done by the author regarding Bayesian Optimization settings in high dimensions.

References

[1] Santu Rana, Cheng Li, Sunil Gupta, Vu Nguyen, Svetha Venkatesh, *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:2883-2891, 2017.