# High Dimensional Bayesian Optimization with Elastic Gaussian Process

**Cheng Li, Santu Rana, Sunil Gupta, Vu Nguyen, Svetha Venkatesh**
Centre of Pattern Recognition and Data Analytic (PraDa), Deakin University
Email: cheng.l@deakin.edu.au

## Abstract

Bayesian optimization depends on solving a global optimization of a acquisition function. However, the acquisition function can be extremely sharp at high dimension - having only a few peaks marooned in a large terrain of almost flat surface. Global optimization algorithms such as DIRECT are infeasible at higher dimensions and gradient-dependent methods cannot move if initialized in the flat terrain. We propose an algorithm that 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. Experiments clearly demonstrate the utility of the proposed method at high dimension using synthetic and real-world case studies.

## 1 Introduction

Bayesian optimization is a sequential procedure where a probabilistic form of the unknown function is maintained using a Gaussian process (GP). A GP is specified by a mean function and a covariance function. A popular choice of covariance function is the squared exponential kernel [8]. A crucial parameter of the kernel is the length-scale which dictates prior belief about the smoothness of the objective function. The posterior of a Gaussian process is analytically tractable and is used to estimate both the mean and the variance of the estimation at unobserved locations. Next, a cheap surrogate function is built that seeks the location where lies the highest possibility of obtaining a higher response. The possibility is expressed through a variety of acquisition functions which trade-off exploitation of the predicted best mean and exploration around high predicted variance.

Acquisition functions are continuous functions, yet they may be extremely sharp functions at higher dimensions, especially when the size of observed data is small. Generally, they have some peaks and a large area of mostly flat surface. For this reason, the global optimization of high-dimensional acquisition functions is hard and can be prohibitively expensive. This makes it difficult to scale Bayesian optimization to high dimensions. Generic global optimization algorithms such as DIRECT [4] perform reasonably when the dimension is low, but at higher dimensions they can become extremely inefficient and actually become infeasible within the practical limitation of resource and time [7]. Multi-start based method start from multiple initializations to achieve local maxima and then choose the best one. However, the multi-start method may not be able to find the non-flat portion of the acquisition function by random search.

Nearly all the existing work assumes that the objective function only depends on a limited number of "active" features [1, 11, 2]. For example, [11] projected the high-dimensional space into a low-dimensional subspace by random embedding and then optimized the acquisition function in a low-dimensional subspace assuming that many dimensions are correlated. This assumption seems too restrictive in real applications [5]. The Add-GP-UCB model [5] allows the objective function to vary along the entire feature domain. The objective function is assumed to be the sum of a set of low-dimensional functions with disjoint feature dimensions. Thus the optimization of acquisition function is performed in the low-dimensional space. [6] further generalized the Add-GP-UCB

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

by eliminating an axis-aligned representation. However, none of them are not applicable if the underlying function does not have assumed structure, that is, if the dimensions are not correlated or if the function is not decomposable in some predefined forms. ***Thus efficient Bayesian optimization for high dimensional functions is still an open problem***.

To address that we propose an efficient algorithm to optimize the acquisition function in high dimension without requiring any assumption on the structure of the underlying function. We recall a key characteristic of the acquisition function that they are mostly flat functions with only a few peaks. Thus gradient-dependent methods would fail to work since a random initialization would most likely fall in the large flat region. However, we theoretically prove that for a location it is possible to find a large enough kernel length-scale which is used to build a new GP and make sure that the derivative of the new acquisition function becomes significant. Next, we theoretically prove that the difference in the acquisition functions is smooth with respect to the change in length-scales, which implies that the extremum of the consecutive acquisition functions are close if the difference in the length-scales is small. We solve a sequence of local optimization problems wherein we begin with a very gross approximation of the function and then the extrema of this approximation is used as the initial point for the optimization of the next approximation which is a little bit finer. Following this sequence we reach to the extrema of the acquisition function for the Gaussian process with the target length-scale which is either pre-specified by the user or estimated for likelihood maximization. We denote our method as **Elastic Gaussian Process** (EGP) method. It is to be noted that our algorithm EGP can easily be converted to pursue a global solution by employing multiple-start with different random initializations.

## 2 Proposed Algorithms

### 2.1 Elastic Gaussian Process

The acquisition function of Bayesian optimization is also associated with the GP kernel length-scale $l$ and hence we denote it as $a(\mathbf{x} \mid \mathcal{D}_{1:t}, l)$. The core task of Bayesian optimization is to find the most promising point $\mathbf{x}_{t+1}$ for the next function evaluation by globally maximizing the acquisition function.

In Lemma 1 we theoretically guarantee that it is possible to find a large enough length-scale for which the gradient of the acquisition function becomes significant at any location in the domain.

**Lemma 1.** $\exists l : \left\| \frac{\partial a(\mathbf{x})}{\partial \mathbf{x}} \right\|_2 \geq \varepsilon$ *for* $l_\tau \leq l \leq l_{max}$*, where* $l_\tau$ *is the target length-scale and* $l_{max}$ *is the maximum length-scale.*

*Proof. The lemma is proved if we prove that* $\left\| \frac{\partial a(\mathbf{x})}{\partial x_i} \right\|_2 \geq \varepsilon, \forall i$*. For UCB[9] after some algebraic manipulation it can be shown that* $\left\| \frac{\partial a(\mathbf{x})}{\partial x_i} \right\|_2 = (\alpha_1/l_2)exp(-\alpha_2/l_2)$*, where* $\alpha_1$ *and* $\alpha_2$ *are constant values. It is easy to show that* $\exists l$ *to satisfy the equality since* $exp(-\alpha_2/l_2)$ *is a decreasing function in* $(0, 1]$ *whist* $\varepsilon l_2/\alpha_1$ *is an increasing function in* $(0, +\infty)$*. Detailed proof has been provided in the supplymentary material. Similar proofs can also be derived for the expected improvement (EI) acquisition function.*

Next in Lemma 2, we theoretically guarantee 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.

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

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

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

*where* $(\mathbf{x}_0, y_0)$ *is the only observation that we have and is used to simplify the proof and* $d_{0i} = x_i - x_{0i}$*.Clearly,* $\frac{\partial g(x_i, l)}{\partial l}$ *is continuous and positive in the domain of* $l$*. Therefore,* $g(x, l)$ *is a smooth function with respect to* $l$*. Again similar proofs can also be derived for the EI acquisition function.*

**Algorithm 1** High Dimensional Bayesian Optimization with Elastic Gaussian Process

```
for $t = 1 : MaxIteration$
    -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 ;
    -evaluate the function $y_{t+1} = f(\mathbf{x}_{t+1})$;
    -augment the data $\mathcal{D}_{1:t+1} = \{\mathcal{D}_{1:t}, \{\mathbf{x}_{t+1}, y_{t+1}\}\}$;
    -update the kernel matrix $\mathbf{K}$;
end for
```

We can now conceive that an algorithm to overcome flat region can be constructed by first finding a large enough length-scale to solve for the optima at that length-scale and then gradually reduce the length-scale and solve a sequence of local optimization problems wherein the optimum of a larger length-scale is used as the initialization for the optimization of the acquisition function based on the Gaussian process with a smaller length-scale. This is continued till the optimum at the target length-scale $l_\tau$ is reached. The whole proposed high-dimensional Bayesian optimization is presented in Alg. 1 and Alg. 2.

## 3 Experiments

We evaluate the EGP on a synthetic example of maximizing a test high-dimensional function and real-world applications of training cascaded classifier. We use preconditioned truncated Newton in NLopt [3]as the local optimization algorithm for each step in our algorithm, which requires gradients to be provided. Our comparators are: 1) A global optimization algorithm using DIRECT (**DIRECT**); 2) Multi-start using preconditioned truncated Newton as a local optimization algorithm (**Multi-start**); 3)High-diemensional Bayesian optimization via additive models [5] (**Add-$d'/M$,** where $d'$ is the dimensionality in each group and $M$ is the number of groups);

Global optimization with DIRECT is used only at lower dimensions of 5 and 10 as it consistently returns with an error at higher dimensions. Multi-start is just the vanilla multi-start without the use of our method. For the additive model, we have multiple instances for different values $(d', M)$. Variables are randomly divided into a set of additive clusters. In all experiments, for fair comparison we use the EI as the acquisition function and the SE kernel as the covariance function. Our algorithms EGP is run with same number of multi-start as the number of dimensions. All the algorithms are given the same fixed time duration. It is computed from the amount of time it takes for our EGP to run. The number of initial observations are set at one more than the number of dimensions. We empirically use the target length-scale $l_\tau = 0.1$ and $l_{max} = 0.9$, $\triangle l = 0.05$. We compare the best evaluation values so far after

**Algorithm 2** Optimizing the acquistion function using EGP

**Input:** a random start point $\mathbf{x}_{init} \in \chi$, $a(\mathbf{x} \mid \mathcal{D}_{1:t}, l)$, the length-scale interval $\triangle l$, $l = l_\tau$.

```
Step 1:
while $l \le l_{max}$ do
    sample $\mathbf{x}^* \leftarrow \text{argmax}_{\mathbf{x}^* \in \mathcal{X}} a(\mathbf{x} \mid \mathcal{D}_{1:t}, l)$ starting with $\mathbf{x}_{init}$;
    if $||\mathbf{x}_{init} - \mathbf{x}^*|| = 0$ then
        $l = l + \triangle l$
    else
        $\mathbf{x}_{init} = \mathbf{x}^*$, break;
    end if
end while

Step 2:
while $l \ge l_\tau$ do
    $l = l - \triangle l$
    sample $\mathbf{x}^* \leftarrow \text{argmax}_{\mathbf{x}^* \in \mathcal{X}} a(\mathbf{x} \mid \mathcal{D}_{1:t}, l)$ starting with $\mathbf{x}_{init}$;
    if $||\mathbf{x}_{init} - \mathbf{x}^*|| = 0$ then
        $\triangle l$=$\triangle l/2$
    else
        $\mathbf{x}_{init} = \mathbf{x}^*$
    end if
end while
```

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

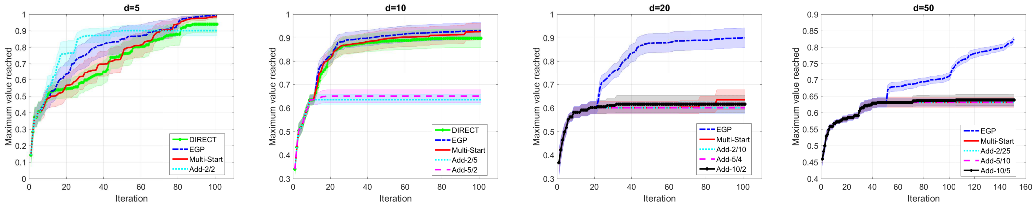

Figure 1: Synthetic example-I: Maximum value reached as a function of Bayesian optimization iteration. Multi-variate unnormalized Gaussian PDF with a maximum of 1 is used as the test function at four different dimensions from left to right (a) d=5 (b) d=10 (c) d=20 (d) d=50. Both mean (line) and the standard errors (shaded region) are reported for 20 trials with random initializations.

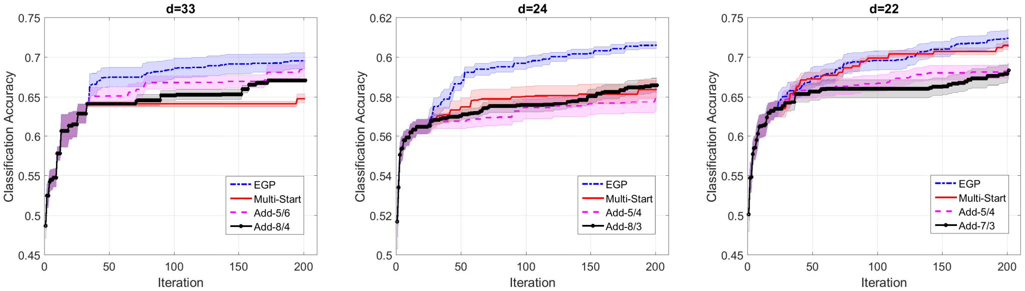

Figure 2: Maximum classification accuracy for training data as a function of Bayesian optimization iteration. We use Adaboost algorithm to learn a cascade classifier. The number of stages is equal to the number of features in three datasets from left to right a) Ionosphere $d = 33$, b) German $d = 24$, c) IJCNN1 $d = 22$. Both mean (line) and the standard errors (shaded region) are reported for 20 trials with random initializations.

each iteration. We run each algorithm 20 trials with different initializations and report the average results and standard errors [1].

### 3.1 Synthetic examples

In this study we demonstrate the application of Bayesian optimization on finding the maximum of an unnormalized Gaussian PDF with a maximum of 1.

We set a block diagonal matrix below as the covariance matrix of the Gaussian PDF $\Sigma = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A} \end{bmatrix}$, where $\mathbf{A} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$. In this case, variables are partially correlated and, therefore, does not admit additive decomposition with high probability when divided randomly. It is further scaled such that the maximum value of the function remains at 1 irrespective of the number of variables (dimensions). Four different experiments have been conducted with four different dimensions 5, 10, 20 and 50. The Bayesian optimization is set-up with a prior Gaussian process that has mean zero function and unit variance. We use the SE kernel with an isotropic length-scale of 0.1. Figure 1 shows the results of Bayesian optimization at four dimensions. Division of variables into additive forms are performed randomly as no information about the function is assumed to be known. We can find that all methods perform similarly at low dimensions ($d = 5$ and $d = 10$). It is because of the low complexity of the acquisition function in the low-dimensional space. Not surprisingly, the EGP performs best at high dimensions ($d = 20$ and $d = 50$) while the additive model fails even at dimension $d = 10$. The EGP does not make any assumptions on the function form or variable correlation.

[1] The code can be downloaded in http://bit.ly/2e5pFDR

### 3.2 Training cascade classifier

Here we evaluate our proposed method by training a cascade classifier [10] on three real datasets from UCI repository: Ionosphere, German and IJCNN1 dataset. A $K$-cascade classifier consists of $K$ stages and each stage has a weak classifier which is a one-level decision stump. Instances are re-weighted after each stage. Generally, independently computing the thresholds are not an optimal strategy and thus in this experiment we wish to find an optimal set of thresholds by maximizing the training accuracy. Features in all datesets are scaled between $[0,\ 1]$. The number of stages is set same with the number of features in the dataset. Therefore, simultaneously optimizing thresholds in multiple stages is a difficult task and thus used as a challenging test case for high-dimensional Bayesian optimization. We create the additive model to make sure the dimension at each group lower than 10 so that DIRECT can work. 200 iterations are performed to optimize the cascade classifier for each dataset. The results are given in Figure 2. For Ionosphere and German datasets, the EGP performs better than other baselines. For IJCNN1 dataset, EGP is slightly better than the Multi-start.

## 4 Conclusion

We propose a novel algorithm for Bayesian optimization in high dimension. At high dimension the acquisition function becomes very flat on a large region of the space rendering gradient-dependent methods to fail at high dimensions. We prove that it is possible to induce a significant gradient at any location in the parameter space by increasing the length-scale of the prior Gaussian process. We further prove that when two Gaussian process priors use only slightly different length-scales then the difference between their acquisition functions remains small. Based on these evidences we formulate our algorithm that first finds a large enough length-scale to enable gradient-dependent methods to perform at any randomly initialized point. After that the length-scale is gradually reduced by using the extremum of the previous acquisition function as the initial point for the next acquisition function. In synthetic and real-world experiments, the proposed method clearly demonstrates very high utility compared to the states of the art at dimensions higher than 10.

## References

[1] Bo Chen, Rui Castro, and Andreas Krause. Joint optimization and variable selection of high-dimensional gaussian processes. In *ICML*, 2012.

[2] Josip Djolonga, Andreas Krause, and Volkan Cevher. High-dimensional gaussian process bandits. In *Advances in Neural Information Processing Systems*, pages 1025–1033, 2013.

[3] Steven G. Johnson. The nlopt nonlinear-optimization package, 2014.

[4] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

[5] Kirthevasan Kandasamy, Jeff G. Schneider, and Barnabás Póczos. High Dimensional Bayesian Optimisation and Bandits via Additive Models. In *ICML*, volume 37, pages 295–304, 2015.

[6] C. Li, K. Kandasamy, B. Poczos, and J. Schneider. High dimensional bayesian optimization via restricted projection pursuit models. In *AISTATS*, pages 1–9, 2016.

[7] Vu Nguyen, Santu Rana, Sunil K Gupta, Cheng Li, and Svetha Venkatesh. Budgeted batch bayesian optimization. In *ICDM*, Spain, 2016.

[8] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[9] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, 2010.

[10] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[11] Ziyu Wang, Masrour Zoghi, Frank Hutter, David Matheson, and Nando De Freitas. Bayesian optimization in high dimensions via random embeddings. In *IJCAI*, pages 1778–1784, 2013.