Accelerator Parameter Optimization using high-dimensional Bayesian Optimization

Project Proposal For Bachelor Thesis

Motivation

Tuning hyperparameters is usually considered a computationally intensive and tedious task, be it for hyper-parameters in neural networks, or complex physical instruments, such as free electron lasers. Users for such applications could benefit from a 'one-click-training' feature, which would find optimal parameters given some reward function in as few steps as possible. This project proposal aims to find such an algorithm which is both efficient, and holds certain convergence guarantees. We focus our efforts in Bayesian Optimization techniques that maximize on a Gaussian Process, and revise techniques for high-dimensional BO.

Background

In Bayesian optimization, we want to use a Gaussian Process to find an optimal parameter setting $\mathbf{x}^*$ that maximizes a given utility function f. We assume the response surface to be Lipschitz-continuous.

Assume we have observations $\mathcal{Y} = \{y^{(1)}, \dots, y^{(N)}\}$, each evaluated at a point $\mathcal{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$. The relationship between the observations y and individual parameter settings $\mathbf{x}$ is $y = f\left(\mathbf{x}\right) + \epsilon$ where $\epsilon \sim \mathcal{N}\left(0, \sigma_n^2\right)$. Any quantity to be predicted has a subscript-star (e.g. y_* is the observation we want to predict).

In it's simplest form, a Gaussian procedure is described by the following equation:

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim N\left(\mu, \begin{pmatrix} K & K_*^T \\ K_* & K_{**} \end{pmatrix}\right), \tag{1}$$

Where $K = \text{kernel}(\mathbf{X}, \mathbf{X})$, $K_* = \text{kernel}(\mathbf{x}_*, \mathbf{X})$ and $K_{**} = \text{kernel}(\mathbf{x}_*, \mathbf{x}_*)$. Using this, the prediction for any point y_*, given all previously sampled points y by estimating the probability $p(y_*|y) \sim N(K_* K^{-1} y, K_{**} - K_* K^{-1} K_*')$

This, in turn, can be used to build an acquisition function. This acquisition function describes where to best sample points next. Some popular acquisition functions include GP-UCB, Most probable improvement (MPI) and Expected Improvement (EI). The choice of the acquisition function has great influence on the performance of the optimization procedure.

We will talk about the problems and possible solutions for the task at hand in the next section.

Scope Of The Project

Bayesian optimization suffers from the curse of dimensionality. The goal of this project is to arrive at a solution that resolves the curse of dimensionality for the specific task at hand with regards to

Bayesian optimization. This project includes, but is not limited to the following methods.

Some proposed solutions are the following:

1. [1] Assume $f(x) \approx g(\mathbf{W}^T x)$ where $\mathbf{W} \in \mathbb{R}^{D \times d}$ and $D >> d$. We can assume that allowing only orthogonal $\mathbf{W}$ is optimal.

   This algorithm does not require gradient-information (thus, easier to implement, and robust to noise). The standard-deviation, kernel parameters and $\mathbf{W}$ can be found iteratively. First we fix $\mathbf{W}$, and optimize over the standard-deviation, kernel parameters. Then we fix the standard-deviation, kernel parameters. and optimize over $\mathbf{W}$. We repeat this procedure until the change of the log-likelihood between iterations is below some ϵ_l.

2. [2] Assume $f(x) = f^{(1)}(x^{(1)}) + f^{(2)}(x^{(2)}) + \ldots + f^{(M)}(x^{(M)})$ where $x^{(i)} \in \mathcal{X}^{(i)} \subseteq \mathcal{X}$, i.e. each function component $f^{(i)}$ takes some lower-dimensional subspace as the input. The lower-dimensional subspaces may overlap. The mean and covariance of $f(x)$ is then the sum of the individual component's means and covariances.

   An additive decomposition (as described above) can be represented by a dependency graph. The dependency graph is built by joining variables i and j with an edge whenever they appear together within some set $x(k)$.

   The goal is to maximize an acquisition function $\phi_t(x) = \sum_{i=1}^{M} \phi_t^{(i)}(x^{(i)})$. This maximization is achieved by maximizing the probability of Markov Random Fields within the graph. A junction tree is created from the graph, which is then used to find the global maximum of the acquisition function.

   The dependencies between the variable-subsets are represented through a graph, which can be learned through Gibbs sampling. This, in turn, is used to create a kernel for the Gaussian process.

3. [3] A function $f : \mathbf{R}^D \rightarrow \mathbf{R}$ is said to have effective dimensionality d_e (where $d_e < D$), if there exists a linear subspace $\mathcal{T}$ of dimension d_e such that for all $x_\top \in \mathcal{T} \subset \mathbf{R}^D$ and $x_\perp \in \mathcal{T}_\perp \subset \mathbf{R}^D$, we have $f(x) = f(x_\top + x_\perp) = f(x_\top)$. $\mathcal{T}^\perp$ is the orthogonal complement of $\mathcal{T}$.

   Assume $f : \mathbf{R}^D \rightarrow \mathbf{R}$ has effective dimensionality d_e. Given a random matrix $\mathbf{A} \in \mathbf{R}^{D \times d}$ (where $d > d_e$) with independent entries sampled by $\mathcal{N}(0,1)$. For any $x \in \mathbf{R}^D$, there exists a $y \in \mathbf{R}^d$ such that $f(x) = f(\mathbf{A}y)$.

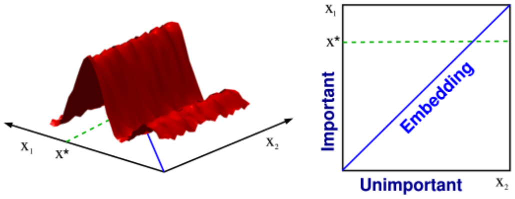

Figure 1: This function in D=2 dimesions only has d=1 effective dimension: the vertical axis indicated with the word important on the right hand side figure. Hence, the 1-dimensional embedding includes the 2-dimensional func- tions optimizer. It is more efficient to search for the opti- mum along the 1-dimensional random embedding than in the original 2-dimensional space

As a result, the optimization task is reduced from a $\mathbf{R}^D$ space to a $\mathbf{R}^d$ space, as we optimize over all possible y (as described in the above definition).

These include currently found methods. If, for some reason, finding an active subspace or an effective lower dimension is not possible, we are open to switch to adapt the procedure of optimization.

References

[1] R. Tripathy, I. Bilionis, and M. Gonzalez, "Gaussian processes with built-in dimensionality reduction: Applications in high-dimensional uncertainty propagation," 2016. [Online]. Available: https://arxiv.org/pdf/1602.04550

[2] P. Rolland, J. Scarlett, I. Bogunovic, and V. Cevher, "High-Dimensional Bayesian Optimization via Additive Models with Overlapping Groups," 2018. [Online]. Available: http://arxiv.org/abs/1802.07028

[3] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. De Freitas, "Bayesian Optimization in High Dimensions via Random Embeddings," 2016. [Online]. Available: https://arxiv.org/pdf/1301.1942