Randomized Algorithms for Gaussian Process Parameter Estimation and Inference

Ross Alexander
Department of Aeronautics & Astronautics
Stanford University
rbalexan@stanford.edu

1 Project Description

Problem Statement In this project, we propose to consider randomized algorithms that can lead to increased performance on parameter estimation and inference in Gaussian processes (GPs) operating on large datasets. In particular, we want to investigate (i) randomized trace estimation methods for producing faster log-likelihood gradient estimates; and (ii) randomized dimension reduction algorithms for projecting large datasets to smaller datasets and therefore improving the overall sample complexity of both parameter estimation and inference procedures. We may consider active learning approaches using GPs, if time permits.

Related Work & Motivation We are interested in this work as GPs are known to have poor sample complexity for inference (for n samples, the computational complexity is $\mathcal{O}(n^3)$ due to kernel matrix inversion). On the other hand, as flexible non-parametric models, GPs have significant expressive power. As a result, significant research has been focused on methods for improving sample complexity. We will conduct a thorough literature review to determine current state-of-the-art methods for inference in GPs (sparse GPs, deep GPs, inducing point methods, etc.).

Methods For our problem, we have a GP given by $\mathcal{GP}(m_\theta, k_\theta)$, where $m_\theta : \mathbb{R} \to \mathbb{R}$ is our mean function and $k_\theta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is our kernel function, both of which are parameterized by θ. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of n input-output pairs (x, y), or *training examples*, we seek to estimate the parameters θ that maximize the (log-)likelihood of the observed data, i.e., we want to solve

$$\max_\theta \mathcal{L}(\theta) = \log p(\mathbf{y} \mid X, \nu, \theta)$$

where ν is a noise variance parameter and $\mathcal{L}(\theta)$ is given by:

$$\mathcal{L}(\theta) = \log p(\mathbf{y} \mid X, \nu, \theta)$$
$$\mathcal{L}(\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det(\mathbf{K}_\theta(X,X) + \nu\mathbf{I}) - \frac{1}{2}(\mathbf{y} - \mathbf{m}_\theta(X))^\top(\mathbf{K}_\theta(X,X) + \nu\mathbf{I})^{-1}(\mathbf{y} - \mathbf{m}_\theta(X)).$$

In order to maximize the log-likelihood, we need the gradient of the log-likelihood, which we compute below for the special case of a zero mean function and zero noise variance.

$$\frac{\partial\mathcal{L}}{\partial\theta_i}(\theta) = \frac{1}{2}\mathbf{y}^\top{\mathbf{K}_\theta}^{-1}\frac{\partial\mathbf{K}_\theta}{\partial\theta_i}{\mathbf{K}_\theta}^{-1}\mathbf{y} - \frac{1}{2}\mathrm{tr}\left({\mathbf{K}_\theta}^{-1}\frac{\partial\mathbf{K}_\theta}{\partial\theta_i}\right) \tag{1}$$

We will attempt to speed gradient computations by performing randomized trace estimation on the trace term. For this proposal, we leave out a more detailed discussion of the projection method, which can be fleshed out in the project.

Datasets & Experiments We will provide a theoretical analysis of each method. For (i), we would potentially like to derive first and second moment expressions for the randomized trace, along with bounds on the gradient approximation error. For (ii) we would like to explore the theory of subspace projection and seek simpler formulae, expressions for (reduced) complexity, and, if possible, expressions for GP first and second moments along with error bounds on the prediction task.

We will test our methods on a variety of large datasets from the UCI Machine Learning Repository[1] and potentially other code-friendly repositories (Kaggle, Scikit-Learn, etc.). We will report the relevant dataset parameters (number of training examples, input and output dimensionality, etc.) as well as runtime with mean and standard deviation.

For gradient estimation, we intend to compute the norm of the trace approximation error and the norm of the composite gradient approximation error. We would like to visualize some of the optimization trajectories between the traditional method and our method to gain a better understanding of the qualitative impact of randomized trace estimation on the likelihood maximization process. For randomized projection, we intend to investigate several transforms (sub-Gaussian, Hadamard, Rademacher, etc.) and report the prediction error (RMSE and ISE over some interval).

Expected Results We expect to see interesting results in applying randomized trace estimation for generating gradient estimates when fitting GPs. We hope these estimates are robust and fast, thereby decreasing the overall time complexity of generating gradient estimates. We may see interesting results for randomized projection.

References

[1] https://archive.ics.uci.edu/ml/datasets