Tutorial: Active Learning for Gaussian Process Regression

Jonas El Gammal

Abstract

This tutorial introduces Active Learning strategies for Gaussian Process (GP) regression. Active Learning aims to reduce the number of training points needed for a model to achieve desired performance by intelligently selecting which data points to evaluate. We discuss the general concept of Active Learning, Bayesian Optimization, and Bayesian Quadrature as specific applications. The material is based on research from my master thesis and PhD thesis on accelerating Bayesian inference using Gaussian processes.

1 Introduction

Active Learning (see e.g., [6] for a review) is an umbrella term for methods that aim to reduce the number of training points needed for a model to achieve a desired performance. In machine learning terminology, the idea is to let the model decide which data points to label next, rather than providing a predetermined dataset.

In physics terms, assume that the goal is to learn a mapping $f(\boldsymbol{x})$ from some input space $\mathcal{X}$ to some output space $\mathcal{Y}$. In passive learning, the model is fed a fixed set of training points $\mathbf{X}, \boldsymbol{y}$ where $\boldsymbol{y} = f(\mathbf{X})$ and trained on this data. In Active Learning, the model is allowed to (typically iteratively) modify the current set of training points to optimize the performance of the model. This is particularly useful when evaluating $f(\boldsymbol{x})$ is computationally expensive or when training on large datasets becomes slow.

Active Learning can be done in a discrete space (when there is only a finite dataset to choose from) or in a continuous space (if $f(\boldsymbol{x})$ can be evaluated at any point). For applications in Bayesian inference of continuous parameters, the latter case is of interest.

Typically, the task of Active Learning strategies is to find the smallest set possible that still allows the model to reach a certain performance. This can be done in conjunction with the task of exploring the sampling space as efficiently as possible.

To illustrate the concept, consider Metropolis-Hastings MCMC. A simple way to extend MH-MCMC to an Active Learning strategy would be to modify the proposal distribution as the sampling progresses, as done in e.g., [4, 3] by using a covariance matrix that is estimated from the current state of the chain.

Active Learning typically relies on some measure to quantify the informativeness of a data point given the current state of the model. This measure can be optimized to find new samples to query, and its choice depends on what the model is supposed to achieve. In the context of GPs, this measure is called the *acquisition function* and is typically a function of the posterior GP.

In the context of GP regression for surrogate inference of probability densities, two concepts are of particular interest: *Bayesian optimization* and *Bayesian quadrature.*

2 Bayesian Optimization

Bayesian optimization (BO) is a method for optimizing black-box functions that are expensive to evaluate. It uses the GP as a surrogate model and iteratively optimizes an acquisition function to determine the next point to evaluate.

2.1 The Expected Improvement Acquisition Function

An example of an acquisition function is the *Expected Improvement* (EI), which is defined as

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}\left[\max(0, f(\boldsymbol{x}) - f(\boldsymbol{x}^+))\right], \tag{1}$$

where $\boldsymbol{x}^+$ is the point with the highest value of f found so far. The goal is to identify the point expected to provide the greatest improvement over the current best point, given the model.

For GPs, the EI is analytical and given by [2]

$$\mathrm{EI}(\boldsymbol{x}) = (f(\boldsymbol{x}) - \mu(\boldsymbol{x}^+))\Phi(z) + \sigma(\boldsymbol{x})\varphi(z), \tag{2}$$

where $z = (f(\boldsymbol{x}) - \mu(\boldsymbol{x}^+))/\sigma(\boldsymbol{x})$ and $\Phi = 1/2\left(1 + \mathrm{erf}(z/\sqrt{2})\right)$ and $\varphi = 1/\sqrt{2\pi}\exp(-z^2/2)$ are the standard normal distribution CDF and PDF, respectively. The point that maximizes the EI is evaluated next. This process is repeated until a stopping criterion is reached.

From the structure of Equation 2, one can see two competing terms: one term is proportional to the difference between the current best point and the point to be evaluated and encourages sampling near the current best point (*exploitation*). The other term is proportional to the uncertainty of the GP and encourages sampling in regions of high uncertainty (*exploration*). This is a common feature in Active Learning strategies and is known as the *exploration-exploitation trade-off*.

2.2 Other Acquisition Functions

The EI is just one of many possible acquisition functions. Alternatives include:

- *Probability of Improvement* (PI) [2]: Measures the probability that a point will improve over the current best.
- *Upper Confidence Bound* (UCB) [1]: Balances exploration and exploitation by adding a multiple of the uncertainty to the mean.

These acquisition functions offer different balances between exploration and exploitation, and the choice depends on the specific application.

3 Bayesian Quadrature

While Bayesian Optimization focuses on optimizing a target function, Bayesian Quadrature (BQ) applies similar principles to estimate the integral of a function [5].

3.1 The Bayesian Quadrature Framework

Let

$$I = \int_{x_0}^{x_1} f(x)\mathrm{d}x \tag{3}$$

be the integral of some arbitrary function $f(\boldsymbol{x})$ over possibly multiple dimensions such that $\boldsymbol{x} \in \mathbb{R}^d$. We restrict ourselves to the one-dimensional case for simplicity, but the generalization to $\boldsymbol{x} \in \mathbb{R}^d$ is straightforward.

The idea is to use the GP as a surrogate model for the integrand. The expectation value of the integral is then

$$\mathbb{E}[I|D] = \int_{x_0}^{x_1} \mu_{f|D}(x)\mathrm{d}x, \tag{4}$$

and the variance

$$\mathrm{var}[I|D] = \int_{x_0}^{x_1} \int_{x_0}^{x_1} \mathrm{cov}_{f|D}(x, x')\mathrm{d}x\mathrm{d}x', \tag{5}$$

where $\mu_{f|D}$ and $\mathrm{cov}_{f|D}$ are the mean and covariance of the posterior GP (conditioned on the training data D), respectively. This not only provides an estimate of the integral but also an estimate of its uncertainty. With this, it is possible to construct an acquisition function that is optimized to find the next point to evaluate.

The disadvantage of this approach is that the d and $2d$ integrals in Equations 4 and 5 still have to be computed numerically. However, for certain kernels (particularly the RBF kernel), these integrals can be computed analytically.

An illustration of BQ is shown in Figure 1, showing how the GP formulation induces a normal distribution over the value of the integral.

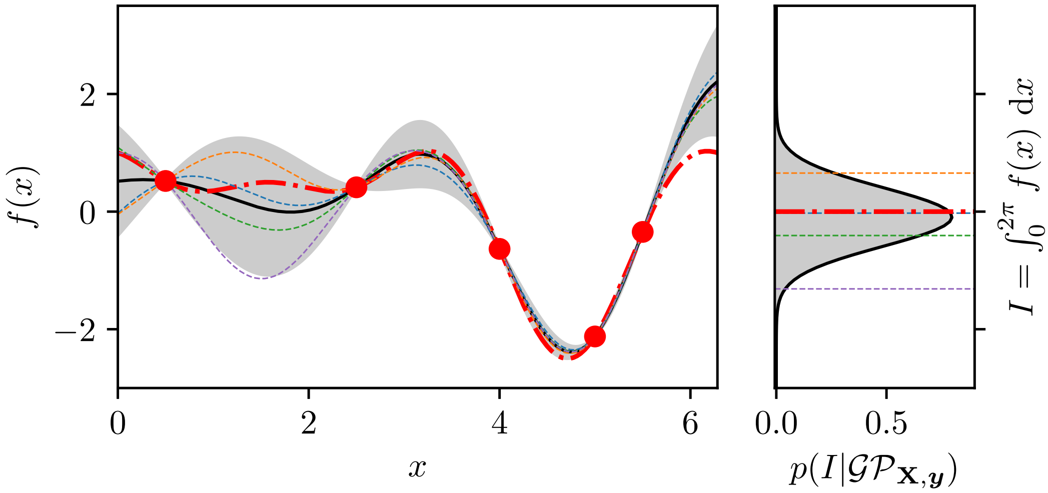

Figure 1: Illustration of the BQ procedure. **Left:** GP fit to the function $f(x) = \sin(x) + \cos(2x) - 1/2\sin(3x)$ (red). The GP has been conditioned on 5 training points (red dots). $\mu(x)$ and $\sigma(x)$ are shown as the black line and gray band, respectively. The dashed lines show four sample functions drawn from $\mathcal{GP}(x|X, y)$. **Right:** Normal distribution induced by BQ for the integral $I = \int_0^{2\pi} f(x)\underset{.}{\mathrm{x}} = 0$ (gray distribution). The true value is shown in red. The four dashed lines correspond to the integrals of the sample functions.

3.2 Advantages of Bayesian Quadrature

Bayesian Quadrature offers several advantages for numerical integration:

- **Uncertainty quantification**: BQ provides not just an estimate of the integral, but also an uncertainty estimate, allowing for principled stopping criteria.

- **Active learning**: By choosing points that maximize the reduction in uncertainty about the integral, BQ can be more sample-efficient than traditional quadrature methods.

- **Adaptive**: The method automatically adapts to the complexity of the integrand, placing more samples where the function varies rapidly.

- **Robustness**: Unlike many traditional methods, BQ can handle irregular or discontinuous integrands.

3.3 Application to Bayesian Inference

In the context of Bayesian inference, BQ can be used to estimate the evidence (marginal likelihood):

$$Z = \int L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{6}$$

where $L(\boldsymbol{\theta})$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is the prior distribution. Traditional Monte Carlo methods like MCMC require many likelihood evaluations ($\gtrsim 10^3$) to correctly recover the shape of the posterior. BQ, on the other hand, can achieve comparable accuracy with significantly fewer evaluations by intelligently choosing where to sample.

The key insight is that BQ treats the integration problem as a regression problem: we want to learn the integrand well enough to estimate its integral. By using Active Learning strategies, we can focus our limited evaluations on the regions that contribute most to the integral or where our uncertainty about the integral is highest.

4 The Exploration-Exploitation Trade-off

A fundamental challenge in Active Learning is balancing *exploration* and *exploitation*:

- **Exploration**: Sampling in regions where the GP has high uncertainty to reduce overall uncertainty about the function.
- **Exploitation**: Sampling in regions where we expect to find good values (for optimization) or high contribution to the integral (for quadrature).

Different acquisition functions provide different trade-offs:

- Pure exploration would sample where $\sigma(\boldsymbol{x})$ is largest.
- Pure exploitation would sample where $\mu(\boldsymbol{x})$ is largest (for maximization).
- Expected Improvement balances both by considering both the mean and the uncertainty.

The optimal balance depends on the specific application and the cost of function evaluations. For expensive likelihood functions in Bayesian inference, a more exploratory strategy early on can pay off by building a better global model before focusing on the high-likelihood regions.

5 Practical Considerations

When implementing Active Learning for GP regression, several practical considerations arise:

5.1 Computational Cost

The main computational costs in GP-based Active Learning are:

1. **GP training**: $\mathcal{O}(N^3)$ for matrix inversion, where N is the number of training points.
2. **Acquisition function optimization**: Requires finding the maximum of the acquisition function, typically done with gradient-based or global optimization methods.
3. **Function evaluation**: The actual evaluation of $f(\boldsymbol{x})$, which may be very expensive for complex simulations.

For Bayesian inference applications, the function evaluation cost typically dominates, making the Active Learning overhead worthwhile if it significantly reduces the number of required evaluations.

5.2 Batch Acquisition

In some applications, it is possible to evaluate multiple points in parallel. In this case, *batch acquisition* strategies can be used to select multiple points simultaneously. One approach is the *Kriging Believer* algorithm, which:

1. Selects the first point by optimizing the acquisition function.
2. Temporarily adds this point to the training set with its predicted mean value.
3. Selects the next point based on the updated GP.
4. Repeats until the desired batch size is reached.
5. Evaluates all points in the batch in parallel.

5.3 Stopping Criteria

Determining when to stop Active Learning is crucial. Possible stopping criteria include:

- **Maximum evaluations**: Stop after a fixed number of function evaluations.
- **Uncertainty threshold**: Stop when the uncertainty about the quantity of interest (e.g., the integral) falls below a threshold.
- **Acquisition function value**: Stop when the acquisition function value becomes small, indicating diminishing returns from additional samples.
- **Convergence**: Stop when the model predictions stabilize between iterations.

For Bayesian inference, a natural stopping criterion is when the uncertainty about the evidence or posterior moments falls below a desired tolerance.

6 Conclusion

This tutorial has introduced Active Learning strategies for Gaussian Process regression, with a focus on:

- The concept of Active Learning and its advantages for expensive function evaluations
- Bayesian Optimization and the Expected Improvement acquisition function
- Bayesian Quadrature for numerical integration with uncertainty quantification
- The exploration-exploitation trade-off in Active Learning
- Practical considerations for implementing these methods

Active Learning with GPs provides a powerful framework for efficiently learning expensive functions and performing numerical integration. By intelligently choosing where to sample, these methods can achieve significant computational savings compared to passive learning or traditional Monte Carlo methods. This makes them particularly valuable for applications in Bayesian inference where likelihood evaluations are computationally expensive.

References

[1] Thomas Desautels, Andreas Krause, and Joel W. Burdick. Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization. *Journal of Machine Learning Research*, 2014.

[2] Donald R. Jones. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, 21(4):345–383, December 2001.

[3] Antony Lewis. Efficient sampling of fast and slow cosmological parameters. *Physical Review D*, 87, 2013.

[4] Antony Lewis and Sarah Bridle. Cosmological parameters from CMB and other data: A Monte Carlo approach. *Physical Review D*, 66, 2002.

[5] Michael Osborne, Roman Garnett, Zoubin Ghahramani, David K Duvenaud, Stephen J Roberts, and Carl Rasmussen. Active Learning of Model Evidence Using Bayesian Quadrature. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[6] Burr Settles. *Active Learning.* Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham, 2012. ISBN 978-3-031-00432-2 978-3-031-01560-1.