GP surrogates can utilise active learning

- Different strategies for selecting parameter values where to query the simulator
  - Reduce the number queries required to produce reasonable approximations to posterior/likelihood
- Usually based on optimization
  - Parameters that produce minimum discrepancies
  - Parameters that decrease most the uncertainty about the posterior

BOLFI

Journal of Machine Learning Research 17 (2016) 1-47 Submitted 1/15; Revised 8/15; Published 8/16

Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models

Michael U. Gutmann MICHAEL.GUTMANN@HELSINKI.FI
Helsinki Institute for Information Technology HIIT
Department of Mathematics and Statistics, University of Helsinki
Department of Information and Computer Science, Aalto University
Jukka Corander JUKKA.CORANDER@HELSINKI.FI
Helsinki Institute for Information Technology HIIT
Department of Mathematics and Statistics, University of Helsinki

Editor: Nando de Freitas

Abstract

Our paper deals with inferring simulator-based statistical models given some observed data. A simulator-based model is a parametrized mechanism which specifies how data are generated. It is thus also referred to as generative model. We assume that only a finite number of parameters are of interest and allow the generative process to be very general; it may be a noisy nonlinear dynamical system with an unrestricted number of hidden variables. This weak assumption is useful for devising realistic models but it renders statistical inference very difficult. The main challenge is the intractability of the likelihood function. Several likelihood-free inference methods have been proposed which share the basic idea of identifying the parameters by finding values for which the discrepancy between simulated and observed data is small. A major obstacle to using these methods is their computational cost. The cost is largely due to the need to repeatedly simulate data sets and the lack of knowledge about how the parameters affect the discrepancy. We propose a strategy which combines probabilistic modeling of the discrepancy with optimization to facilitate likelihood-free inference. The strategy is implemented using Bayesian optimization and is shown to accelerate the inference through a reduction in the number of required simulations by several orders of magnitude.

Keywords: intractable likelihood, latent variables, Bayesian inference, approximate Bayesian computation, computational efficiency

1. Introduction

We consider the statistical inference of a finite number of parameters of interest $\boldsymbol{\theta} \in \mathbb{R}^d$ of a simulator-based statistical model for observed data $\mathbf{y}_o$ which consist of n possibly dependent data points. A simulator-based statistical model is a parametrized stochastic data generating mechanism. Formally, it is a family of probability density functions (pdfs) $\{p_{\mathbf{y}|\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$ of unknown analytical form which allow for exact sampling of data $\mathbf{y}_{\boldsymbol{\theta}} \sim p_{\mathbf{y}|\boldsymbol{\theta}}$. In practical terms, it is a computer program which takes a value of $\boldsymbol{\theta}$ and a state of the random number generator as input and returns data $\mathbf{y}_{\boldsymbol{\theta}}$ as output. Simulator-based models are also called implicit models because the pdf of $\mathbf{y}_{\boldsymbol{\theta}}$ is not specified explicitly (Diggle and Gratton, 1984), or generative models because they specify how data are generated.