EPSRC CDT in Distributed Algorithms

PhD Project: Developing Efficient Numerical Algorithms Using Fast Bayesian Random Forests

University of Liverpool

PhD Student: Harvinder Lehal
Project Partner: GCHQ

Supervisors:
Dr. Navjot Kukreja, University of Liverpool
Dr. Lee Devlin, University of Liverpool
GCHQ

Project Description

Decision Trees are used in data science to estimate parameter values to classify data sets in terms of regression and discrete labels. Current attempts to estimate the parameters use methods like Bayesian Additive Regression Trees (BART) and Classification Additive Regression Trees (CART), relying on Bayesian models, such as Markov Chain Monte Carlo (MCMC), which are arguably less sophisticated than what can be achieved by using Random Forests. MCMC, despite having generalised variants such as No-U-Turn-Samplers (NUTS), is hard to use in applications where the number of parameters is often unknown, making it unusable for looking at trees where the number of nodes, branches and thus parameters changes every iteration.

Instead of looking at individual trees, a collection of trees, a Random Forest can be used. They are in pervasive use in data science and machine learning. In such algorithms, each tree describes succinct rules that relate the inputs to the outputs (which can be both continuous values, in the context of regression, and discrete labels, in the context of classification). An advantage of using random forests is the ability to account for missing or lost data values and the ability to look at many more trees at once.

This PhD will seek to apply efficient, numerical Bayesian algorithms, such as Sequential Monte Carlo (SMC) Samplers, that can be used to develop a novel variant of a Random Forest algorithms. The intent is that the new algorithms would be drop-in replacements to the Random Forest, but can use modern parallel computational resources such as GPUs to do parallel programming, to use a given training dataset to provide more accurate predictions in the same elapsed time compared to the MCMC samplers used for individual trees.

Go to the EPSRC CDT In Distributed Algorithms website.