Part III Project Plan

Drawing Samples from probability distributions in multi-dimensions: improving on a poor state-of-the-art

Expectations of variables, or functions of variables, under a probability distribution may be found by using a sample of random data points from within the distribution. However, if the probability distribution is not adequately approximated by a Gaussian – for example if there is more than one peak in the distribution (multimodal) – sampling randomly from the distribution is not a trivial problem, especially in high-dimensional spaces.

The problem is challenging because without prior knowledge of the distribution one cannot know whether much of the probability is contained in unboundedly-high spikes in certain locations in the parameter space. Correct samples from the distribution will by definition come from these places where the probability is high, and so the method used must discover them, and find out how much of the total probability is contained in them.

Clever Monte-Carlo type methods have been developed to solve these problems – though as the title suggests, the state-of-the-art algorithms have numerous flaws – and this project will be looking at these. During the Michaelmas Term 2009, the relevant literature has been consulted, and work has begun on the computational implementation of a sampling method.

One such method, developed in Cambridge by Allanach and Lester[1] for use in tens of dimensions to investigate Beyond-the-Standard-Model physics, is called 'bank sampling'. This method is an implementation of the Metropolis-Hastings Algorithm (well described in Mackay's book[2]), and uses a 'bank' of parameter space points which act as clues for areas of interest in the distribution.

The main goal of this project is to implement a bank sampler and investigate its strengths and weaknesses. This will be done by developing 'toy problems' to probe the known difficulties and observed limitations of the algorithm, and to investigate its behaviour in high-dimensional spaces. It is hoped that it will be possible to develop 'pathologically good' and 'pathologically bad' distributions, which demonstrate situations where the algorithm works well and fails respectively. From this understanding, it might become possible to develop improvements to the algorithm in order to reduce or remove its present problems, or at least make some suggestions in this direction.

Whilst the project will mainly focus on the bank sampler, there is another algorithm with the same motivation, also developed at Cambridge by Feroz and Hobson[3], and known as 'multimodal nested sampling'. Whereas the Metropolis-Hastings algorithm, of which bank sampling is a special case, makes use of a Markov chain to 'randomly' walk around the parameter space, the nested sampling approach uses Monte-Carlo random points within ellipsoidal nests which are placed around the probability distribution, thus confining the search area to more interesting regions.

It is hoped that in the course of the project, if time permits, it might be possible to look at multimodal nested sampling in a similar way to bank sampling, in order to draw comparisons between them, and hopefully develop an understanding of where each is good and bad. It might then be possible to develop heuristics for how to choose between these two methods for a given multi-dimensional problem.

Dr C.G. Lester

[1] http://arxiv.org/abs/0705.0486

[2] D.J.C. Mackay, *Information Theory, Inference and Learning Algorithms*, CUP 2003, pp. 365–370

[3] MultiNest: http://www.mrao.cam.ac.uk/software/multinest/ and http://xxx.lanl.gov/abs/0704.3704v3