Response to Referee Comments: Parallel Adaptive Importance Sampling

February 3, 2017

First of all, we would like to thank the referees for their time in reading and commenting on our manuscript. In this document we will address each of the points raised, and describe any relevant changes that have been made to the paper. The changes made have been substantial, in particular with the addition of a 5-dimensional multimodal example arising from mixture models.

Referee 1 comments

Comment: *My strong suspicion is that the scheme is convergent when both the number of copies (N) and the number of MCMC iterations goes to infinity ... but this isn't shown here..*

Answer: We thank the referee for this comment. We have added a new section in the manuscript which points to proofs of both of the types of convergence that the referee alludes to.

Comment: *It seems likely that the paper was distilled from a Ph.D. thesis ... it's a bit longer than it has to be and includes some boilerplate background material. But other than that it's well written..*

Answer: The paper was certainly lengthy, and as such we have distilled it down somewhat. We have added a new example, but have removed the biochemical network example, which required lengthy introduction, and one of the earlier 1-dimensional examples. The introduction has been completely rewritten, and includes more material which is directly applicable to the paper. We hope that these extensive changes have made for a better paper with a slicker presentation. Even with other additions, such as the section on consistency, we have reduced the length of the paper to a leaner 24 pages.

Comment: *The examples are really basic and seem aimed to describe implementation rather than to convince us that the method will work. On more* *complicated problems I'd be concerned that the weights would be very degenerate and the resampling step would collapse all the walkers on top of each other ... in that case the different chains would be very highly correlated and you'd be doing a lot of work for no reason..*

Answer: We agree that if the proposal kernels are chosen poorly or with a poor choice of parameters, then problems can exist. All poor importance sampling schemes will result in the behaviour described. It is important that the proposal kernels are absolutely continuous with respect to the posterior distribution, and it is important that the variance of the kernels is appropriately chosen such that the mixture proposal distribution is slightly overdispersed. This has been further discussed and highlighted in the new submission.

We concede that the examples presented were relatively simple. We have removed a couple of these, and have included a new example arising from mixture models. The posterior in the new example is 5-dimensional, and has two modes which are well separated. Metropolis-Hasting algorithms really struggle in these scenarios, since the waiting time for the chains to switch modes is very large. Since many transitions are required for convergence, the convergence of these methods is glacial. We demonstrate the PAIS's aptitude for this type of problem with the new example, which we believe greatly improves the paper.

Comment: *Even if that work is done in parallel you have to compare to the same number of independent chains..*

Answer: As we hope is clear, all the comparisons made between Metropolis-Hastings and their PAIS equivalents are made with the same number of chains in both. We are not comparing our method with one serial chain of MH, as is common in most of the other parallel MCMC literature. What stands out with this method is the fact that we get superlinear improvement in the number of iterations that are required for convergence with respect to the number of chains used.

Referee 2 comments

Comment: *The work in [1] on parallel samplers for climate model parameter estimation is definitely relevant here, but has not been cited. Same goes for the adaptive MCMC work of Haario, e.g. [2]..*

Answer: We thank the referee for these suggestions which we have now included

Comment: *The "pre-fetching" way of parallelizing samplers (where the possible future acceptances/rejections are calculated in advance) would also be worth mentioning here..*

Answer: We have added a reference to this approach in the introduction.

Comment: *The numerical examples are very simple (only 1d and 2d), and somewhat fail to convince me that this is a good approach..*

Answer: As described above in the response to the first referee, a new 5-dimensional multimodal example has been added in which MH algorithms really struggle. We feel this demonstrates the significant advantages of this method.

Comment: *How would this scale to higher dimensional problems? Finding a good importance function is hard in high dimensions..*

Answer: This approach, like all importance sampling schemes, suffers from the curse of dimensionality. As the dimension increases, the number of particles which are required in order to achieve a good proposal distribution increases fast, and as such high-dimensional problems are beyond the scope of this work. We have rewritten the introduction to emphasise that this approach is more suited to low-dimensional challenging posterior distributions for which Metropolis-Hastings methods fail to converge quickly. This is demonstrated in our new 5-dimensional example as described above.

Comment: *The adaptation only aims at learning the scale of the proposal, whereas the shape can be as (or more) important..*

Answer: We partly disagree with this comment, since the positions of the chains themselves give us the shape of the posterior distribution. This said, maybe the referee was referring to the shape of the individual proposal kernels which together make up the mixture proposal distribution. In the situation where the posterior distribution is highly concentrated on a lower dimensional manifold, as is ubiquitous in many applications, such as inverse problems for biochemical networks and epidemiology, then if the shape of this correlation is not learned by the algorithm, then the result of the standard PAIS approach may not be optimal. We are currently working on a new paper where we use PAIS to sample from a transformation of the posterior distribution. This transformation is an approximation of the transformation which maps the posterior to a standardised N-dimensional Gaussian, which naturally is easy to sample from. This is an extensive piece of work, and beyond the scope of this paper.