Response to Referee Comments: Parallel Adaptive Importance Sampling

December 13, 2018

First of all, we would like to thank the referees and the associate editor 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. We have made several substantial changes, including the renaming of the algorithm and paper from "parallel adaptive importance sampling" (PAIS) to "ensemble transport adaptive importance sampling" (ETAIS). We have also replaced three of the old numerical examples with two ones, one a bimodal example, and the second a more challenging problem for Bayesian inversion of the Lorenz '63 equations. We believe that the paper is stronger, and thank the reviewers and associate editor for their input.

Referee 1 comments

Comment: *This is supposed to be a parallel algorithm and yet it's implementation is serial. Furthermore, the authors don't seem to be entirely sure that parallelization will work since they comment that it's effectiveness will depend on various factors. The title seems somehow misleading to me. Maybe if the "parallel" part of the title is removed and the discussion about parallelization is treated as a possibility instead of a critical part of the algorithm, then the document will be less confusing..*

Answer: Despite the fact that we felt that we had demonstrated the feasibility of the parallelisation of this method, we agree with the referee that we have not demonstrated this in the numerics. Furthermore, this algorithm delivers a speed-up even when run in serial. As such, we are happy to change the name of the algorithm to "Ensemble Transport Adaptive Importance Sampling" (ETAIS). This required some rewriting throughout the paper. In particular, we have toned down the emphasis on parallelisation, and put more emphasis on the fact that this algorithm also delivers speed-up for serial implementaions.

Comment: *Also, the comparison between serial PAIS and the naive parallel*

MCMC is strange if there is no parallelization..

Answer: The naive parallel chains are run for long enough that they are statistically equivalent to one long chain. We have added a discussion on this point to *****

Comment: *It will be interesting that the authors show the PAIS using AMS with the examples. They show the difference between the AMS and the ETPF but not within the PAIS..*

Answer: We assume that the referee meant the AMR? As the AE asked for a new example to replace the second example, we took this as an opportunity to compare ETAIS with the ETPF, AMR and bootstrap resamplers, which is now presented in the new first example. The numerics demonstrate that superior samplers lead to more stable and accurate sampling, with ETPF algorithm narrowly outperforming the AMR algorithm, with the bootstrap version following up in the rear.

Comment: *There is an error at the end of page 24..*

Answer: Many thanks for spotting this, this has now been corrected.*******

Referee 2 comments

Comment: *The paper is ready for publication..*

Answer: Many thanks. We hope you agree that the additional changes have further improved the paper.

Associate Editor comments

Comment: *The authors have made a substantial improvement over previous version of the paper. However, there are a number of points that I would like to see clarified before the paper is accepted for publication.*
1) The posterior sometimes is called μ sometimes π some times $\pi(x|D)$ etc. In fact, there is a confusion in Alg. 1 in p.6 with this..

Answer: We have carefully gone through the paper and clarified where appropriate whether we are referring the posterior measure μ or its density π, and unified the notation as much as possible.

Comment: *2) Choosing the proposal (kernel) in a MH is all the main and*

commonly the only problem in MCMC. PAIS also needs a proposal of exactly the same kind. Why can we expect that within PAIS this will be easier? An indeed, PAIS will perform better that a naively parallelized MH, both using the same proposal. How PAIS saves some of the problems in finding good proposals?.

Answer: As discussed in the paper, there are many possible choices for the proposal kernel around each of the ensemble members. The only thing that we require to ensure convergence of an AIS algorithm, is the absolute continuity of the kernel with respect to the posterior. However, to ensure stability of the algorithm, we also require that the tails of the proposal distributions are at least as fat as the target. In practice, this means that picking kernels which are of the same type as the prior is usually a good idea. The ETAIS methodology does not, however, greatly benefit from the use of more informed and expensive proposals, such as are seen in the standard M-H framework, where gradient information can be used to improve proposals (MALA, HMC, etc), since the information about the whole target is represented by the ensemble. As mentioned in the paper, we have conducted similar studies with MALA proposals, and although we sometimes observe a slightly quicker burn-in period, overall the performance is largely the similar to that of a RW proposal, at twice the cost. We have added a detailed discussion around this to ******

Comment: *3) It might be trivially to do so, but the authors only briefly mention a prove for the convergence of their algorithm. In fact, is it ok to use a dependent proposal in importance sampling? ... no regularity conditions are needed? Why Alg. 3 works? resampling "introduces error" ... is this relevant? does the error cascades down in each iteration? In the examples the PAIS seem to work ok, but, do we need any regularity condition? Please make your convergence claims more explicit and take more time to explain them. In particular, do we have the correct invariant distribution?.*

Answer: This algorithm belongs to a family of methods for which convergence has been proven in the literature, and as such, we have referenced this proof. We have clarified this a little further in *******

Comment: *4) There are too many examples. The univariate gaussian is not very informative and the univariate bimodal, is not either since both modes are of the same scale. I would rather see these two examples turned into one single one with a 2D mixture of two separated gaussians with contrasting scales (variance covariance matrices)..*

Answer: We agree that the numerical experiments section is very long. However the examples presented do each demonstrate a different aspect of the advantages of our approach. We have taken this comment on board, however, and have drastically reduced the length of example 1, in which we demonstrate that tuning the proposal variances using the effective sample size is approximately equivalent to using the L^2 errors in the target, which is only available to

us as we have picked a trivial target. We have replaced the second bimodal example, with a bimodal example as requested by the AE with differing scales in the covariances of the two modes. We have also taken this opportunity to demonstrate the differences in results when using the different resamplers discussed in the paper. We have also included a comparison with a standard bootstrap resampler. The new numerics demonstrate the improvement in stability and convergence of the method with the use of more advanced resampling methods. The mixture model example demonstrates the algorithm working in a higher dimensional multimodal setting. The final new example (see later comments) demonstrates the speed-ups achieved using our approach for a challenging problem with an expensive likelihood, requiring the approximation of a differential equation. We believe that each example serves a purpose, and we have also reduced the size of this section by ***** pages.

Comment: *5) In the examples the authors include very ad hoc strategies like the 'scout' chain and specially the very ad hoc strategy described in section 7.4.2 (!). The authors, in my opinion, do not really discuss the implementation of the PAIS algorithm but a case-by-case ad hoc version of it. This is clearly unacceptable..*

Answer: The scout chains were only used in the old bimodal example, and have since been cut. This approach has already been used in the pMC literature to stabilise the importance sampling by fattening the tails of the proposal distribution, but we have no problem with removing it from this paper.

As for the proposal kernels used in 7.4.2, we disagree that this approach is "ad-hoc". These are chosen simply because the proposal distribution must only be supported by [0,1]. We have gone into more details regarding this and explaining the choices, and how users might make their own choices for different scenarios, in *********.

Comment: *6) Please clarify the first phrase of section 7.4.2, which, as far as I see, is simply twice wrong. There is a (very simple!) analytic version of the posterior AND with the right proposal MH can sample from it (!)..*

Answer: We have amended the statement about the analytic form of the posterior. We are not aware of any M-H methods which can reliably sample from multimodal densities with well-separated modes. HMC trajectories can be hand-tuned to make transitions between modes, but only with information regarding the separation. Ensemble-based methods appear to us to be the only candidate for reliable sampling from such distributions without such a hand-tuning. However, we have agreed to remove the claim.

Comment: *7) Although the chaotic Lorenz 93 ODE is used, it is used in such a short time scale that the posterior is quite well behaved (unimodal and fairly gaussian and only some correlation is seen between to parameters). Since*

the priors are normal, this regime suggest a linear part of the FM. The authors should attempt something more challenging, by running the system into a highly nonlinear regime. Otherwise, why consider this example?.

Answer: The reason for considering the example was that the AE had previously asked us to produce an example in the regime where we suggest that we get the biggest win using our proposed method, i.e. those where the likelihood evaluations have significant cost due to requirement of the solution of a differential equation. In this example, we have demonstrated a significant speed-up over the equivalent M-H algorithm. We do understand however the AE's desire for a more challenging inverse problem in itself, and so we have created a new example which has a bigger prior distribution. The push-forward of this prior distribution is multimodal on the timescale of the observations (see the figure below), showing that trajectories do pass close to a hyperbolic point in the assimilation window. This also results in a longer thinner posterior, which as can be seen from the numerics, RWMH really struggles to sample from efficiently. The ETAIS implementation handles this problem with ease, for a big speed-up.

Comment: *8) Do not use an explicit Euler solver, use any higher order, off the shelve solver..*

Answer: Many of the off-the-shelf higher order solvers are just as inappropriate to use for the approximation of the solution of the Lorenz '63 equations, since they are chaotic, and as such *any* numerical error is quickly amplified through the dynamics. Moreover, the focus of this paper is not on these applications, but in the statistical algorithm. In this example, we are simply looking for a non-linear mapping such that we can demonstrate the speed-up that is gained from using ETAIS as opposed to standard M-H sampling, and we believe that a simple Euler approximation provides such a mapping.