AISTATS 2017

Artificial Intelligence and Statistics 2017

Apr 20, 2017 - Apr 22, 2017, Fort Lauderdale, Florida, USA

Reviews For Paper

Paper ID	339
Title	An Efficient Minibatch Acceptance Test for Metropolis-Hastings

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Summary of Paper and Detailed Comments	This paper introduces an approximate minibatch acceptance test for Metropolis-Hastings. In the big N regime, Metropolis-Hastings suffers the most from the computation of acceptance probabilities where likelihoods need to be computed over the entire dataset. The proposal in this work is to consider the log likelihood over a minibatch of data as approximately Gaussian and add a correcting random variable to derive an acceptance test derived from a Barker test function. Comments: - Regarding the requirement that the variance of the log acceptance probability is bounded. The authors claim in the introduction that reducing the step size is one possible solution. Yet, this seems to conflict with the last few sentences of section 2.2 where the authors claim that this methods opens up the possibility of more aggressive step sizes. Clarification would be valuable here. - The bounds on the stationary distribution are appreciated. - The empirical results seem impressive.
Overall Rating	Weak Accept
Reviewer Confidence	Understood the main points in the paper, but skipped the proofs and technical details, not very familiar with the literature

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Summary of Paper and Detailed Comments	Summary of the paper The paper presents a subsampling-based MCMC algorithm, for problems where the number of items in the dataset is intractably large. Summary of the review In the spirit of [Korattikara et al. 2014, Bardenet et al. 2014], the authors propose a test-based acceptance step using Barker's kernel and a Berry-Esseen-type inequality. The topic is definitely of interest, but my major concern is that a lot of probabilistic arguments are shaky and need to be cleaned. Comments p = page c = column rv = random variable - p2 c2 "which for general distributions requires..." In [Bardenet et al. 2014, 2015], the authors give examples of models for which the bounds only require sweeping once over the dataset, as a preprocessing. - Eqn (12) you seem to assume the two rv's are independent, as supported by the convolution in Eqn (14), but I don't understand why they would be.

	- Eqn (14): I'm not sure I understand why we want C_\sigma to satisfy (14). - Lemma 3: what is "an approximate Student's distribution"? In particular, you have not assumed the rv's are Gaussian, do you rely on some limit argument? There are more precise probabilistic tools for this, such as Berry-Esseen's inequality. - Lemma 5 is only applicable if you know there exists a stationary distribution to the approximate kernel, so you need to prove there is one for your algorithm. - Section 5: there is no clear theorem that states the error of your algorithm. Interesting tools are presented, but connections are left for the reader to make. - Section 6.1: could you use the control variate trick of [Bardenet et al. 2015] to reduce the cost per iteration of their algorithm? Minor comments p2 c1 "MH can be used in similar fashion to SGD": I think this statement is too vague, and potentially misleading. Eqn (5) I don't see the batch size b defined anywhere before this equation. Eqn (7) isn't the '1' in psi supposed to be a 'u'?
Overall Rating	Reject
Reviewer Confidence	Understood the paper. Checked the proofs. Familiar with the literature

Masked Reviewer ID: Assigned_Reviewer_4

Review:

Question	
	This paper proposes a novel approach to run approximate M-H acceptance test efficiently. The main idea is to decompose the logistic random variables in the Barker algorithm into a Normal random variable and a correction random variables. Then the error in the mini-batch based estimate of the log-likelihood ratio is consumed approximately as part of the Normal random variables. Unlike previous works in Korattikara et al. [2014] and Bardenet et al. [2014] where the approximation error depends on the variance of the likelihood estimation noise, the approximation error in this paper depends on the accuracy of the correction distribution, given the variance of the estimation noise is small enough and satisfies the CLT. This paper shows that the confidence level does not have to be discounted and the size of data usage per iteraion are much smaller than Korattikara et al. [2014] and Bardenet et al. [2014]. This is a quite novel approach in approximate M-H algorithms and the algorithm is well explained and examined in this paper. The paper is well organized and easy to follow. The propsed algorithm has an intrinsic difficulty that I'll explain later, but I think it still makes a good contribution to approximate M-H algorithms and introduces new perspective to battle with the noise introduced from data subsampling. The proposed method has an intrinsic shortcoming that is the variance of the likelihood ratio has to be smaller than 1. The authors argue that this can be overcome by either increasing the minibatch size or increasing the temperature. However, for the former solution, for some problem with large variance in the log-likelihood ratio the minibatch size has to be increased to a very large value and thereby it loses the advantage over existing methods. For the latter solution, increasing temperature changes the problem of interest and is not desirable in most cases. I was wondering if it is possible to combine the proposed algorithm with sequential test based algorithms such as Korattikara et al. [2014]. Since the two algorithms are the same except for the termination rule and the accept/rejection rule, can we choose an algorithm based on what condition is satisfied at every iteration. If the noise variance is small enough, one can use the proposed algorithm, and if the

Summary of Paper and Detailed Comments	variance is large but much smaller than the mean, one can use the rule in Korattikara et al. [2014] to terminate the sequential test. The bounds on the stationary distribution still applies with this hybrid approach. Another remedy is to adopt the variance reduction technique used in section 3.4 of Chen & Ghahramani (2016). Unlike the method in Bardenet et al. [2015], this method does not reply on an assumption of approximate Normal posterior distribution, and it reduces the variance significantly without changing the target distribution. The authors also provide a similar but improved version of Korattikara et al. [2014] that takes into account the error discounting. I have a question about Table 1. In the text of section 4, “Table 1 shows that the errors between Xnorm+Xcorr and Xlog approach single floating precision (about 10e-7)” but Table 1 actually shows the L_\inf error is >= 5e-6. Also, the values of N, sigma, and lambda in Table 1 are chosen without explanation. As the authors use sigma=1 in the algorithm, what’s the error in that case and how the error depends on N and lambda? Besides, should one get a better solution to u by replacing the L2 regularisation with linear constraints u_i >= 0? It can be solved efficiently with quadratic programming. The following comments are about experiments. In both experiments, the authors use a high temperature in order to run the proposed algorithm. It is OK for proof of concept, but it does not make sense for a practical use such as the logistic regression experiment for MNIST. Why do we compare the proposed algorithms at a very high temperature while the competing algorithms can run at the original distribution? Also, one should get a much faster convergence rate by replacing the random proposal with the SGLD proposed and combine it with an approximate acceptance test as in Korattikara et al. [2014]. In section 6.1 and 6.2, the authors mention “Thus our results for Korattikara et al. [2014] should be treated as lower bounds.” and “so again our results represent a lower bound for this method.” It is unclear to me what it means by “lower bound” here. In the logistic regression experiment, the proposed method achieves a similar convergence rate as Korattikara et al. [2014] measured by cumulative data usage but it uses much less data per iteration. Does it mean the proposed algorithm makes more errors per iteration? Reference: Yutian Chen, and Zoubin Ghahramani. "Scalable Discrete Sampling as a Multi-Armed Bandit Problem." Proceedings of The 33rd International Conference on Machine Learning, pp. 2492–2501, 2016
Overall Rating	Accept
Reviewer Confidence	Understood the paper. Checked the proofs. Familiar with the literature

Masked Reviewer ID: Assigned_Reviewer_5

Review:

Question	
	The paper aims to develop a more efficient, approximate subsampling-based MCMC algorithm based around the approximation of the acceptance probability within Barker's algorithm using estimates obtained from minibatches. It presents some analysis which attempts to bound the error induced in terms of quantities which can be estimated as well as an application to one toy example and a subset of the MNIST classification problem.

Summary of Paper and Detailed Comments	The use of subsampling-based techniques to reduce the computational cost of MCMC algorithms in the large data regime is certainly topical and has attracted a good deal of attention in recent years. This paper proposes a modification to existing methods which it is argued improves computational performance quite substantial and allows the approximation error to be bounded using available quantities. However, the contribution is somewhat incremental and leaves a number of questions unanswered. I am not entirely convinced that the error bounds provided are quite as much better than those available for existing schemes as claimed. The quantities involved in the bounds of Bardenet et al. may not be readily calculable in all models, but they are bounds. The use of an asymptotic bound, with quantities within it estimated using small minibatches which it is conjectured will be adequate yields an estimate, but not a bound and so I find the paragraph following corollary 1 at best rather optimistic. See. Bardenet et al. Section 4.1 for a discussion of bounds based upon asymptotic arguments. The use of tempering with the paper is not adequately explained and I find it slightly worrying that it seems to be suggested in a number of places that tempering can be employed by simply conducting "inference at a different temperature from that assumed by data generation". In Section 6.2 for example is written "The temperature is set at T=1000" without explanation or any further comment. It's commonly argued that simulation-based Bayesian inference is justified in spite of the greater cost associated with the approach than many schemes for point inference because of its intrinsic characterization of uncertainty, but this benefit seems to be lost if tempering is used to simplify the problem without some subsequent correction. Readers would want to know exactly what is done in this respect. There seem to be some approximations within the algorithm which are not studied or discussed: The method seems to appeal to asymptotic normality of \lambda^* which would seem to require that b (i.e. the subsample size) is large enough to justify this asymptotic approximation. How is this assessed? The scaling of X_nc in the algorithm uses the sample variance as though it were the actual variance, which seems potentially rather dangerous given that minibatches may be rather small. How is the resulting error assessed / controlled? Notation is a little confusing in places: * Using $x_1,...,x_b$ to refer to the b observations in a particular minibatch is confusing as they differ from the first b of the n observations denoted in the same way. * What exactly is intended by saying that \Lambda_i is an iid rv on page 3? Conditional on the observed xs, if the \Lambda_i arises from sampling an observation uniformly at random then it's a conditionally iid RV, but if \Lambda_i is deterministically related to x_i in the deterministic way described then this breaks down. The paragraph following lemma 3 is rather unclear. It's not possible for \bar{X} to converge to a distribution whose "variance is measured from the sample". The result shows the difference between the distribution of \bar{X} and a normal with a variance depending on the sample converges to zero; provided that the sample variance itself converges then the distribution of \bar{X} converges to a normal with that variance. Results for the GMM example naively seem to show that the approach of Bardenet et al. better approximates the distribution than the method described here. This is dismissed because "the variance of these values is high and ordering changes depending on the range of samples generated" and while I'm not quite sure what is

	meant by "the range of samples generated" this seems to suggest that these methods have not been run for long enough to characterize the statistics being reported. Results from the MNIST example, the tempering notwithstanding, do show a certain cost regime in which the method seems to outperform its competitors but the description in text seems to focus rather heavily on the positive aspects of the graphs shown being based essentially on behaviour between 10^5.5 an 10^6 processed data. Small queries: p2, col2, "the probability of a transition..." unless this is a discrete space, this isn't a probability. p2 and later: it is repeatedly emphasized that unbiased estimates of log likelihoods are used. Is this unbiasedness relevant? The paper is enthusiastic about subsampling in HMC, both in Section 2.2 and in future work; how is this enthusiasm reconciled with the content of: Betancourt, Michael. "The Fundamental Incompatibility of Scalable Hamiltonian Monte Carlo and Naive Data Subsampling." Proceedings of The 32nd International Conference on Machine Learning. 2015.? p4, "the objective can be written as:" is the expression that follows not an approximation of the objective function given previously based around discretisation? The phrasing seems to suggest that it is exactly the same.
Overall Rating	Reject
Reviewer Confidence	Understood the paper. Checked the proofs. Familiar with the literature