NIPS 2016

Neural Information Processing Systems 2016

December 05 - 10, 2016, Barcelona, Spain

Reviews For Paper

Paper ID	1035
Title	A Simple Minibatch Acceptance Test for MCMC

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Summary: Provide a brief summary of the contents of the paper.	The paper addresses MCMC for tall data problems, proposing a variant of the acceptance step of Metropolis-Hastings that only considers a subsample of the full dataset.
Fatal flaws: Does the paper have a "fatal flaw" making it unfit for publication, regardless of other criteria (may include out of scope, double publication, plagiarism, wrong proofs, flawed experiments)? Use the text box to justify your answer.	No (not as far as I can see)
Technical quality: whether experimental methods are appropriate, proofs are sound, results are well analyzed.	2-Sub-standard for NIPS
Novelty/originality: in any aspect of the work, theory, algorithm, applications, experimental.	3-Poster level (some notable novel contributions)
Potential impact or usefulness: could be societal, academic, or practical and should be lasting in time, affecting a large number of people and/or bridge the gap between multiple disciplines.	3-Poster level (looks promising)

Clarity and presentation: explanations, language and grammar, figures, graphs, tables, proper references.	2-Sub-standard for NIPS

Qualitative assessment: Provide constructive feedback to the authors; justify and complement your ratings above. This is our MOST IMPORTANT QUESTION, we need to get good arguments for our decisions!	# Summary of the review
	The topic of MCMC for big datasets is of importance, and the idea of swapping subsampling noise for acceptance noise is interesting, although not new, see below. My main concern is that paper is not clear nor technically sound enough for publication in NIPS. In particular, a more standard and detailed analysis is needed, see major comments below for hints.
	
	# Major comments
	[BDH] = [Bardenet, Doucet and Holmes, MCMC for tall data, http://arxiv.org/abs/1505.02827].
	- Section 2: there are a lot of uncited works, see e.g. the recent survey [BDH].
	- L80 "hurt mixing": this is too vague, what does it hurt? Is geometric ergodicity not preserved, for example? Or does it increase asymptotic variance while preserving the type of ergodicity?
	- Lemma 2 is too vague a statement to be a lemma. What does "approximately Gaussian" mean? Also, in the proof, a CLT for sampling without replacement is used, this requires a reference as it is nonstandard. Or is it because N is large that sampling without replacement is approximated with sampling with replacement?
	- L132 "the noise means": this is unclear.
	- Lemma 2: [BDH] warn against using CLT-based approximations in Metropolis-Hastings and give examples where the resulting algorithms perform worse than simple SGD. This should be commented. In particular, it feels a bit weird to assume n is large enough for a CLT (per MCMC iteration!) to hold, while we are trying to keep n low.
	- Section 3.2: the fact that part of the subsampling noise is transferred to the acceptance noise in Eqn (7) is interesting. This trick is not completely new though, see [BDH, Section 6.3] for instance.
	- The decomposition in Eqn (6) is not unique, and is thus ill-defined.
	- L136: X_norm and epsilon are both random variables, so I wouldn't use the term "estimate".
	- Overall, Section 4 is confusing and should be rephrased. For instance, using P both for kernels and probability of accceptance is confusing. Same for notations such as a circle for the action of kernels on distributions.
	- L201: who are the $\pi'_i$?
	- It is hard for the reader to draw a conclusion from Section 4. Traditionally, this analysis section should first focus on whether the considered approximate algorithm has a limiting distribution, and then try to show a law of large numbers or a CLT, see e.g. [Douc, Moulines, Stoffer, Nonlinear time series, 2014, Sections 5-7]. This would be more informative.
	
	# Minor comments
	- L63 "accepting if" -> "if and only if".
	- L285 a lot of references are lacking capitals.

Reviewer confidence regarding this review.	3-Expert (read the paper in detail, know the area, quite certain of my opinion)

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	

Summary: Provide a brief summary of the contents of the paper.	This paper proposes a new approach to carrying out mini-batch based acceptance test for MCMC. It introduces a novel and interesting idea of absorbing the noise of the log-probability ratio estimate by decomposing the random variable X into a normal distribution and a correction term. Unfortunately, the proposed acceptance function mixes slower than the Metropolis acceptance function, the theoretical contribution is not rigorously proved, and the experiment setting is problematic and the result do not show convincing evidence that the proposed algorithm is better than the adaptive M-H test.
Fatal flaws: Does the paper have a "fatal flaw" making it unfit for publication, regardless of other criteria (may include out of scope, double publication, plagiarism, wrong proofs, flawed experiments)? Use the text box to justify your answer.	Yes
Explain fatal flaws if you ticked yes in the previous question.	The theoretical contribution is not rigorously proved. The experiment setting is problematic and the result do not show convincing evidence that the proposed algorithm is better than the adaptive M-H test. Please see my detailed comments below.
Technical quality: whether experimental methods are appropriate, proofs are sound, results are well analyzed.	1-Low or very low
Novelty/originality: in any aspect of the work, theory, algorithm, applications, experimental.	4-Oral level (significantly novel and impressive)
Potential impact or usefulness: could be societal, academic, or practical and should be lasting in time, affecting a large number of people and/or bridge the gap between multiple disciplines.	3-Poster level (looks promising)
Clarity and presentation: explanations, language and grammar, figures,	2-Sub-standard for NIPS

graphs, tables, proper references.	
Qualitative assessment: Provide constructive feedback to the authors; justify and complement your ratings above. This is our MOST IMPORTANT QUESTION, we need to get good arguments for our decisions!	Summary: This paper proposes a new approach to carrying out mini-batch based acceptance test for MCMC. It introduces a novel and interesting idea of absorbing the noise of the log-probability ratio estimate by decomposing the random variable X into a normal distribution and a correction term. Unfortunately, the proposed acceptance function mixes slower than the Metropolis acceptance function, the theoretical contribution is not rigorously proved, and the experiments do not show convincing evidence that the proposed algorithm is better than the adaptive M-H test. Qualitative assessment: This paper proposes a novel approach of conducting M-H test based on a mini-batch. I really like the idea of decomposing the distribution of the random X and use the normal part to absorb the error in the estimation of the log-probability ratio. This idea is quite original and could make a good impact if properly adopted. Unfortunately, this paper does not convert this idea into a practical algorithm. The algorithm design, theoretical analysis and empirical evaluations are problematic. This paper first proposes a different acceptance function from the Metropolis function. The proposed M-H test with the logistic function is actually the Barker algorithm, first introduced in (Barker, 1965). I understand that the authors choose the logistic distribution in order to get a symmetric random distribution, but the Barker function is less efficient than the original Metropolis-Hastings algorithm. E.g., when the proposed location \theta' is as good as \theta, M-H accepts it with a probability 1 but the Barker function has rejects it with a probability of 0.5. Problem with the decomposition of X: There is not an accurate definition of X_{norm} and X_{corr}. There are infinitely many ways to decompose X in Equation 6. It's not clearly what it means by "If X_{norm} is exact" in line 136. Although we do not instantiate X_{norm} in the algorithm, in the theoretical analysis in Sec 4, we need to a constructive definition for X_{norm} and X_{core} in order to study their property. E.g. in line 179, when we define X_{\xi}, how do we compute X_{norm} in the first hand? The theoretical analysis of Sec 4 is not rigorous. - In Lemma 2, central limit theorem applies for sufficiently large N and n and n << N. - In the proof of Lemma 3 and 4, we should use the mean value theorem to prove the first inequality in each proof rather than the Taylor expansion (high order term is not guaranteed to be smaller in all cases). - The upper bound in Lemma 3 should be $2 \zeta l$ rather than $2 \zeta l$ because the total variation distance is 0.5 * l_1 distance. - Line 197, the maximum value of X_{\xi} is unbounded if distributed as Gaussian. Also, by definition, the variance of X_{\xi} is not necessarily smaller than \epsilon. - What conclusion can get draw from Theorem 1? Is it possible for the difference of the approximate Markov chain from the exact posterior to diverge? - The analysis of the approximation Markov chain does not consider the case when the estimated standard deviation exceeds 1.2. Simply skipping the iteration will break the detailed balance. Problems with the experiment setting: First of all, I do not agree that increasing the temperature is the correct way of satisfying the variance precondition. We should not avoid a weakness of an algorithm by changing the original problem we want to solve. Reducing the step size of the proposal distribution is a more reasonable choice. Second, could the author explain how to choose the hyper-parameters of the mini-batch algorithm and the adaptive test in the experiment?

	In experiment 5.1, it is hard to tell if the proposed method is better than the adaptive algorithm. Please show the two figures in Fig. 3 both in log-count or linear count for a fair comparison. Does the proposed algorithm include the K mini batches that are used to estimate std at every iteration into the count? I think the best way to compare the two algorithm is to estimate the effective sample size of each algorithm as a function of the number of processed data or running time, while the bias of the approximate distribution is controlled at the same level. In experiment 5.2, according to the appendix, the proposed algorithm draws samples in the distribution with temperature = 3000 while the adaptive MH algorithm runs in the original distribution. We should not compare the accuracy or log-likelihood of two algorithms in two different posterior distributions. In experiment 5.3, dropout is usually used for deep models during training but not used in the test phase. It is hard to provide a Bayesian explanation for dropout in this Bayesian NN problem. Also, the proposed algorithm does not help improve the log-likelihood. Reference: Barker, A.A.: Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. Australian Journal of Physics 18, 119–133 (1965)
Reviewer confidence regarding this review.	3-Expert (read the paper in detail, know the area, quite certain of my opinion)

Masked Reviewer ID: Assigned_Reviewer_4

Review:

Question	
Summary: Provide a brief summary of the contents of the paper.	The article proposes a mini-batch acceptance test for Markov Chain Monte Carlo algorithm. Compared to Balan et al. (2014), the proposed algorithm uses a fixed mini-batch size instead of an adaptive one. The author illustrated the performance via several examples.
Fatal flaws: Does the paper have a "fatal flaw" making it unfit for publication, regardless of other criteria (may include out of scope, double publication, plagiarism, wrong proofs, flawed experiments)? Use the text box to justify your answer.	No (not as far as I can see)
Technical quality: whether experimental methods are appropriate, proofs are sound, results are well analyzed.	2-Sub-standard for NIPS
Novelty/originality:	

in any aspect of the work, theory, algorithm, applications, experimental.	2-Sub-standard for NIPS
Potential impact or usefulness: could be societal, academic, or practical and should be lasting in time, affecting a large number of people and/or bridge the gap between multiple disciplines.	2-Sub-standard for NIPS
Clarity and presentation: explanations, language and grammar, figures, graphs, tables, proper references.	3-Poster level (good enough)
Qualitative assessment: Provide constructive feedback to the authors; justify and complement your ratings above. This is our MOST IMPORTANT QUESTION, we need to get good arguments for our decisions!	First, the discussion of using the logistic acceptance function in Page 3 seems quite off topic. With the original MH test, the acceptance rule still takes the form of $\Delta > \log u$ where u ~ uniform(0, 1). The logistic acceptance function actually has a uniformly lower acceptance rate than the original MH test. Correct me if I was wrong, but I don't see the advantage of using logistic acceptance test here. Second, the central limit theorem only guarantee $\sqrt{N} \sum_i \log p(x_i\|\theta)$ converge to Gaussian. It does not guarantee $N \sum_i \log p(x_i\|\theta)$ converge to any distribution. Also I didn't quite get the argument in the whole Line 131 - 154. Third, the theoretical analysis is mostly within the framework of Balan et al. (2014), i.e., the original MCMC has to be uniformly ergodic and so on. However, the conditions are kind of weird. For example, for Lemma 3, it requires the estimator X_\xi to be uniformly bounded, which is almost impossible for distributions with unbounded support. In addition, the requirement on the bounded jumping step is also weird. It is almost impossible to control the jumping step in the actual implementation. Thus, the analysis is not very convincing.
Reviewer confidence regarding this review.	2-Confident (read it all; understood it all reasonably well)

Masked Reviewer ID: Assigned_Reviewer_6

Review:

Question	
Summary: Provide a brief summary of the contents of the	In MCMC methods, the acceptance test requires computing the likelihood of all instances in the training set. In this paper, the authors present a method for approximating the acceptance probability using a small number of subsets of the training data sampled at random (minibatches). They present theoretical

paper.	guarantees on convergence of the new method and empirically demonstrate significantly improved mixing times as a function of the computation time.
Fatal flaws: Does the paper have a "fatal flaw" making it unfit for publication, regardless of other criteria (may include out of scope, double publication, plagiarism, wrong proofs, flawed experiments)? Use the text box to justify your answer.	No (not as far as I can see)
Technical quality: whether experimental methods are appropriate, proofs are sound, results are well analyzed.	4-Oral level (top 3% submissions)
Novelty/originality: in any aspect of the work, theory, algorithm, applications, experimental.	3-Poster level (some notable novel contributions)
Potential impact or usefulness: could be societal, academic, or practical and should be lasting in time, affecting a large number of people and/or bridge the gap between multiple disciplines.	4-Oral level (many people will pick this up)
Clarity and presentation: explanations, language and grammar, figures, graphs, tables, proper references.	4-Oral level (excellent in every respect)
Qualitative assessment: Provide constructive feedback to the authors; justify and complement your ratings above. This	Good potential for impact. There seems to be significant interest in this problem, and the authors' model seems very easy to deploy in practice, while providing good theoretical guarantees on performance with only a reasonable set of basic assumptions. The authors compare with a good selection of current approaches to this problem and demonstrate significant advantages to their model. The authors also do a good job identifying and justifying the important motivating

is our MOST IMPORTANT QUESTION, we need to get good arguments for our decisions!	rationale behind their design and addressing potential drawbacks/considerations for their method. Overall, I think this paper is well positioned, easily digestible, and has good potential for impact.
Reviewer confidence regarding this review.	2-Confident (read it all; understood it all reasonably well)

Masked Reviewer ID: Assigned_Reviewer_7

Review:

Question	
Summary: Provide a brief summary of the contents of the paper.	The paper propose a novel approach to perform Metropolis-Hastings step in Markov Chain Monte Carlo efficiently using a subset of data. Compared with prior work like [Balan et al 2014], it provide a more elaborate strategy. Theoretical results are satisfactory.
Fatal flaws: Does the paper have a "fatal flaw" making it unfit for publication, regardless of other criteria (may include out of scope, double publication, plagiarism, wrong proofs, flawed experiments)? Use the text box to justify your answer.	Yes
Explain fatal flaws if you ticked yes in the previous question.	The author mentions that in this method the size of mini-batch is fixed. However, in Section 3.3, in Variance Preconditioning , the author mentions that " (b) increase the minibatch", contradict?
Technical quality: whether experimental methods are appropriate, proofs are sound, results are well analyzed.	2-Sub-standard for NIPS
Novelty/originality: in any aspect of the work, theory, algorithm, applications, experimental.	2-Sub-standard for NIPS

Potential impact or usefulness: could be societal, academic, or practical and should be lasting in time, affecting a large number of people and/or bridge the gap between multiple disciplines.	2-Sub-standard for NIPS
Clarity and presentation: explanations, language and grammar, figures, graphs, tables, proper references.	4-Oral level (excellent in every respect)
Qualitative assessment: Provide constructive feedback to the authors; justify and complement your ratings above. This is our MOST IMPORTANT QUESTION, we need to get good arguments for our decisions!	This paper develop an alternative minibatch acceptance test for MH step in MCMC method which uses only a subset of data. It is easy to follow and theoretical analysis is thorough. However, it seems that in variance preconditioning, size of minibatch is not fixed, contradicting with the statement in Introduction. Moreover, the authors claim that this method works better than other method under GPU but does not support their view in experiment. Furthermore, in NN experiment, why not compare the proposed method with conventional MH method ? Why adaptive gradient descent method is chosen as baseline method instead of SGD ? The size of dataset also seems limited. It would be more convincing if the author can compare this method with adaptive MH [Balan et al 2014] in more high-dimensional large-scale dataset.
Reviewer confidence regarding this review.	3-Expert (read the paper in detail, know the area, quite certain of my opinion)