View Author Feedback For Paper

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

	Question	Response
1	Rebuttal	**We thank reviewers for their thoughtful comments.** **To specifics:** **First, our method does not require temperature > 1. We reran the logistic regression example at temperature 1 and reduced step size. Generating 3000 samples took 30 secs for our method. Korattikara et al. was >= 10x slower, and Bardenet et al. was 100x slower on the same data. They consumed proportionate data. Convergence is slower because of reduced step size, however we added an annealing phase at T=100 reaching equilibrium at 1500 steps, and ran at T=1 thereafter. In hindsight, using two elevated temperature examples in our submission was a poor choice.** **It is impossible in general to take large steps (i.e. generate uncorrelated samples) on small minibatches of data since there is not enough information to localize the distribution's mode. One can either (i) use data blocks almost as large as the original dataset as in previous work or (ii) increase temperature or decrease step size. We present a method that generates efficient small-minibatch samples assuming (ii), i.e. whenever it is possible to do so. Earlier works fail to generate samples with small minibatches even given (ii).** **(ii) allows many more samples per pass over the dataset, and is far more useful to us as practitioners than (i). We have many results in progress now that use this test for posterior model inference on challenging learning tasks including deep network model inference and distribution optimization.** **Reduced step size is characteristic of other minibatch-based MCMC methods, such as Gibbs sampling, [1] and the papers that build on it. There model parameters derive from global counts of states which are updated incrementally. The effective step size is O(1/N) for an N-point dataset.** **Other papers cited by us: Chen et al. and Welling et al. use vanishingly small steps. Our approach allows substantial step increase before hitting the variance 1 limit (R2). We can list many other papers that highlight community interest in (ii).** **On incremental contribution: We request R5 make a critical review of earlier work. We tried to use it and hit extreme challenges. Bardenet et al (1) requires a global per-sample**

bound. (R3: carefully chosen examples aside, there are no general bounds). A practitioner faced with a large dataset and non-trivial model faces a brick wall.

For Korattikara et al., there are no error bounds. The paper (appendix A) instead assumes "...when the size of a mini-batch is large enough, e.g. n > 100, the central limit theorem applies, and also slj is an accurate estimate of the population standard deviation." It further assumes covariances are exactly measured on minibatches, that minibatch errors are exactly gaussian. It uses a complex dynamic programming method to allow large per-test errors which is sketched in the paper. Details of quantization etc were missing, code was never released and we were not able to replicate it. Hence we describe lower bounds for this method using no test discounting at all (Bardenet et al. use the same simplification in their implementation of Korattikara et al. and released code for it)

By contrast our method requires no global dataset bounds, and uses only moment estimates from minibatches. It is the only "black box" method at this time. i.e. that can generate samples on small batches of data *when possible* while providing accurate error bounds (without per-output-sample statistics from the full data). And it is very simple to implement. We feel this is a very significant contribution.

Details:

(R4) Our analysis is based on the CDF of a t-statistic (sorry for student's distribution comment R3), we can use the empirical sdev as the distribution sdev without separate analysis. i.e.

sup_v|Pr(x/s < v) - Psi(v)| = sup|Pr(x < u) - Psi(u/s)| where u = s * v

and s is the sample sdev. (R3) Most Berry-Esseen bounds require knowledge of the exact sdev, and give the same asymptotic error without explicit constants. They are not useful here.

(R3) Eqn 12, Xcorr is a machine-generated PSN, so necessarily indept.

(R3) Will discuss equilibirum distribution existence.

Yes, (R4) we agree that Chen & Ghahramani (2016) would be a useful (necessarily small) improvement, and will add that to the paper.

(R5) Compared to Bardenet, our method has a "zero difference" accept rate of 0.5 vs 1. It generates more repeated samples (with correct distribution) to which the Chi-squared comparison is sensitive, and appears to do worse.

(R5) We do not appeal to asymptotics nor absolute dataset bounds as in prior work. We use bounds with explicit constants based on moments of the minibatch, which can be estimated quite accurately in practice.

We regret we cannot address the remaining comments in space available. We read all carefully and while the paper is not perfect, more detailed reading of it should address those comments.

		[1] Finding scientific topics: Thomas L. Griffiths and Mark Steyvers
2	confidential comments to Senior Program Committee	

✉ For Website related issues email CMT Support