An Efficient Minibatch Acceptance Test for Metropolis-Hastings*

Daniel Seita[1], **Xinlei Pan**[1], **Haoyu Chen**[1], **John Canny**[1,2]
[1] University of California, Berkeley, CA
[2] Google Research, Mountain View, CA
{seita,xinleipan,haoyuchen,canny}@berkeley.edu

Abstract

We present a novel Metropolis-Hastings method for large datasets that uses small expected-size minibatches of data. Previous work on reducing the cost of Metropolis-Hastings tests yields only constant factor reductions versus using the full dataset for each sample. Here we present a method that can be tuned to provide arbitrarily small batch sizes, by adjusting either proposal step size or temperature. Our test uses the noise-tolerant Barker acceptance test with a novel additive correction variable. The resulting test has similar cost to a normal SGD update. Our experiments demonstrate several order-of-magnitude speedups over previous work.

1 Introduction

Markov chain Monte Carlo (MCMC) sampling is a powerful method for computation on intractable distributions. We are interested in large dataset applications, where the goal is to sample a posterior distribution $p(\theta|x_1, \ldots, x_N)$ of parameter θ for large N. The Metropolis-Hastings method (M-H) generates sample candidates from a proposal distribution q which is in general different from the target distribution p, and decides whether to accept or reject based on an acceptance test [Metropolis *et al.*, 1953; Hastings, 1970].

Many state-of-the-art machine learning methods are based on minibatch updates (such as SGD) to a model. Such updates produce many improvements to the model for each pass over the dataset and have high sample efficiency. In contrast, conventional M-H requires calculations over the full dataset to produce a new sample. Results from [Korattikara *et al.*, 2014; Bardenet *et al.*, 2014] perform approximate (bounded error) acceptance tests using data subsets, but may still require close to the full data. By contrast, [Maclaurin and Adams, 2014] perform exact tests but require a lower bound on the parameter distribution across its domain, and such bounds are only available for relatively simple distributions.

Here we derive a new test which incorporates the variability in minibatch statistics as *a natural part of the test* and requires less data per iteration than prior work. We use a Barker test function [Barker, 1965], an idea which was suggested but not explored empirically in Section 6.3 of [Bardenet *et al.*, 2017]. But the asymptotic test statistic CDF and the Barker function are different, which leads to fixed errors for the approach in [Bardenet *et al.*, 2017]. Here, we show that the difference between the distributions can be corrected with an additive random variable. This leads to a test which is fast, and whose error can be made arbitrarily small.

We note that this approach is fundamentally different from prior work. It makes no assumptions about the form of, and requires no global bounds on the posterior parameter distribution. It is exact in the limit as batch size increases by the Central Limit Theorem. This is not true of [Korattikara *et al.*, 2014; Bardenet *et al.*, 2014] which use tail bounds and provide only approximate tests even with arbitrarily large batches; we instead use moment estimates from the data to determine how far the minibatch posteriors deviate from a normal distribution. These bounds carry through to the overall test accuracy.

Our test is applicable when the variance over data samples of the log probability ratio between the proposal and the current state is small enough (less than 1), a natural condition for models running M-H sampling with optimal proposals [Roberts and Rosenthal, 2001] on a full dataset. It succinctly captures the condition that the minibatch carries enough information to generate a sample. Though we cannot generally expect to get independent samples from the posterior using only a small subset of the data, we can still exploit minibatches by doing the following:

1. Increase the temperature K of the target distribution. Log likelihoods scale as $1/K$, and so the variance of the likelihood ratio will vary as $1/K^2$.
2. For continuous distributions, reduce the proposal step size (i.e. generate correlated samples), which is characteristic of Gibbs samplers on large datasets [Dupuy and Bach, 2016]. The variance of the log acceptance probability scales as the square of proposal step size.
3. Utilize Hamiltonian Dynamics [Neal, 2010] for proposals and tests. Here the dynamics itself provide shaping to the posterior distribution, and the M-H test is only needed to correct quantization error.

In sum, the primary contribution of our work is a novel, more efficient (in samples per test) minibatch acceptance test

*This paper is an abridged version of a paper titled "An Efficient Minibatch Acceptance Test for Metropolis-Hastings" that won Honorable Mention for Best Student Paper award at UAI 2017.

with quantifiable error bounds. The test uses a novel additive correction variable to implement a Barker test based on minibatch mean and variance. We demonstrate several order-of-magnitude improvements in sample efficiency, and that the batch size distribution is short-tailed.

2 Preliminaries

2.1 M-H Background and Related Work

In the Metropolis-Hastings method [Gilks and Spiegelhalter, 1996; Brooks *et al.*, 2011], a difficult-to-compute probability distribution $p(\theta)$ is sampled using a Markov chain $\theta_1, \ldots, \theta_T$. The sample θ_{t+1} at time $t+1$ is generated using a candidate θ' from a (simpler) proposal distribution $q(\theta'|\theta_t)$, filtered by an acceptance test. The acceptance test is usually a Metropolis test. The Metropolis test has acceptance probability:

$$\alpha(\theta_t, \theta') = \frac{p(\theta')q(\theta_t|\theta')}{p(\theta_t)q(\theta'|\theta_t)} \wedge 1 \tag{1}$$

where $a \wedge b$ denotes $\min(a, b)$. With probability $\alpha(\theta_t, \theta')$, we accept θ' and set $\theta_{t+1} = \theta'$, otherwise set $\theta_{t+1} = \theta_t$. The test is often implemented with an auxiliary random variable $u \sim \mathcal{U}(0, 1)$ with a comparison $u < \alpha(\theta_t, \theta')$; here, $\mathcal{U}(a, b)$ denotes the uniform distribution on the interval $[a, b]$. For simplicity, we drop the subscript t for the current sample θ_t and denote it as θ. For Bayesian inference, the target distribution is $p(\theta|x_1, \ldots, x_N)$ with acceptance probability

$$\alpha(\theta, \theta') = \frac{p_0(\theta') \prod_{i=1}^N p(x_i|\theta')q(\theta|\theta')}{p_0(\theta) \prod_{i=1}^N p(x_i|\theta)q(\theta'|\theta)} \wedge 1 \tag{2}$$

where $p_0(\theta)$ is the prior.

Computing samples this way requires all N data points which is intractable for large datasets. To address this challenge, [Korattikara *et al.*, 2014; Bardenet *et al.*, 2014] perform approximate M-H tests using sequential hypothesis testing; we refer to these as AUSTEREMH and MHSUBLHD. At each iteration, a subset of data is sampled and used to test whether to accept θ' using an approximation to $\alpha(\theta, \theta')$. If the approximate test does not yield a decision, the minibatch size increases and the test repeats. Both show useful computational savings but have two drawbacks: (i) they are approximate, and always yield a decision with a finite error, (ii) they require exact, dataset-wide bounds that depend on θ.

There is a separate line of MCMC work based on statistical physics using Langevin Dynamics [Wang and Uhlenbeck, 1945] and Hamiltonian Monte Carlo (HMC) [Neal, 2010; Betancourt, 2017] which can generate distant, high-quality proposals when run on full datasets. Recent work has attempted to combine these methods with SGD-like updates via SGLD [Welling and Teh, 2011; Ahn *et al.*, 2012] and SGHMC [Chen *et al.*, 2014]. Our method can be integrated with these to generate full posterior samples for almost the same cost as SGD.

2.2 Notation

Following [Bardenet *et al.*, 2014], we write the test $u < \alpha(\theta, \theta')$ equivalently as $\Lambda(\theta, \theta') > \psi(u, \theta, \theta')$, where

$$\Lambda(\theta, \theta') = \sum_{i=1}^N \log \frac{p(x_i|\theta')}{p(x_i|\theta)} \tag{3}$$

and $\psi(u, \theta, \theta') = \log\left(u \frac{q(\theta'|\theta)p_0(\theta)}{q(\theta|\theta')p_0(\theta')}\right)$. To simplify notation, we assume that temperature $K = 1$ (saving T to indicate the number of samples to draw). Temperature appears as an exponential on each likelihood, $p(x_i|\theta)^{1/K}$, so the effect would be to act as a $1/K$ factor on $\Lambda(\theta, \theta')$.

To reduce computational effort, an unbiased estimate of $\Lambda(\theta, \theta')$ based on a minibatch $\{x_1^*, \ldots, x_b^*\}$ can be used:

$$\Lambda^*(\theta, \theta') = \frac{N}{b} \sum_{i=1}^b \log \frac{p(x_i^*|\theta')}{p(x_i^*|\theta)}. \tag{4}$$

Finally, define $\Lambda_i(\theta, \theta') = N \log(\frac{p(x_i|\theta')}{p(x_i|\theta)})$, which are i.i.d. random variables. By the Central Limit Theorem, we expect $\Lambda^*(\theta, \theta')$ to be approximately Normal. The acceptance test then becomes a test of the hypothesis that $\Lambda(\theta, \theta') > \psi(u, \theta, \theta')$ by establishing that $\Lambda^*(\theta, \theta') \gg \psi(u, \theta, \theta')$.

3 A New MH Acceptance Test

For our new M-H test, we denote the exact and approximate log likelihood ratios as Δ and Δ^*, respectively. First, Δ is

$$\Delta(\theta, \theta') = \log \frac{p_0(\theta') \prod_{i=1}^N p(x_i|\theta')q(\theta|\theta')}{p_0(\theta) \prod_{i=1}^N p(x_i|\theta)q(\theta'|\theta)}, \tag{5}$$

where p_0, p, and q are the same as in Equation (2). We separate out terms dependent and independent of the data and then write our minibatch estimator Δ^* for Δ as

$$\Delta^*(\theta, \theta') = \underbrace{\frac{N}{b} \sum_{i=1}^b \log \frac{p(x_i^*|\theta')}{p(x_i^*|\theta)}}_{\Lambda^*(\theta, \theta')} - \psi(1, \theta, \theta'). \tag{6}$$

Note that Δ and Δ^* are evaluated on the full dataset and a minibatch of size b respectively. The term N/b means $\Delta^*(\theta, \theta')$ is an unbiased estimator of $\Delta(\theta, \theta')$.

3.1 Barker (Logistic) Acceptance Function

We consider functions other than the classical Metropolis test that satisfy the detailed balance condition needed for accurate posterior estimation, and invoke the following [Barker, 1965]:

Lemma 1. *If $g(s)$ is any function such that $g(s) = \exp(s)g(-s)$, then the acceptance function $\alpha(\theta, \theta') \triangleq g(\Delta(\theta, \theta'))$ satisfies detailed balance.*

For our test we use the logistic function $g(s) = (1 + \exp(-s))^{-1}$, which satisfies Lemma 1. To understand this choice, a test function $f(x)$ for M-H must satisfy Lemma 1. In addition, it must be monotone, bounded by $[0, 1]$ and be such that $\lim_{x\to-\infty} f(x) = 0$ and $\lim_{x\to\infty} f(x) = 1$. The Logistic is the unique function in this class which is anti-symmetric about 0.5, representing the CDF of a symmetric random variable. Our method requires approximating this with the sum of a Normal random variable (also symmetric) and a correction (Section 3.3). The Logistic CDF L and Normal CDF Φ are extremely close even without correction; the CDF error from the closest Normal CDF satisfies $\sup_x |L(x) - \Phi(x/1.7)| < 0.01$. With our correction we can make this error orders of magnitude smaller.

Assume we have current and candidate samples θ and θ', and have $V \sim \mathcal{U}(0,1)$. We accept θ' if $g(\Delta(\theta,\theta')) > V$, and reject otherwise. Since $g(s)$ is monotonically increasing, its inverse $g^{-1}(s)$ is well-defined and unique. So equivalently, we accept θ' if and only if $\Delta(\theta,\theta') > X = g^{-1}(V)$ where X is a random variable with the logistic function as its CDF. The density of X is symmetric, so we equivalently test

$$\Delta(\theta,\theta') + X > 0 \tag{7}$$

for a logistic random variable X.

3.2 A Minibatch Acceptance Test

We now describe acceptance testing using the minibatch estimator $\Delta^*(\theta,\theta')$. From Equation (6), $\Delta^*(\theta,\theta')$ can be represented as a constant term plus the mean of b IID terms $\Lambda_i(\theta,\theta')$ of the form $N \log \frac{p(x_i^*|\theta')}{p(x_i^*|\theta)}$. As b increases, $\Delta^*(\theta,\theta')$ therefore has a distribution which approaches Normal by the CLT, with mean Δ. In the limit, we can write

$$\Delta^* = \Delta + X_{\text{norm}}, \quad X_{\text{norm}} \sim \bar{\mathcal{N}}(0, \sigma^2(\Delta^*)), \tag{8}$$

where $\bar{\mathcal{N}}(0,\sigma^2(\Delta^*))$ denotes a distribution which is approximately normal with variance $\sigma^2(\Delta^*)$. But to perform the test in Equation (7) we want $\Delta + X$ for a logistic random variable X (call it $X_{\log}$ from now on). In [Bardenet *et al.*, 2017] it was proposed to use Δ^* in a Barker test, and tolerate the fixed error between the logistic and normal distributions. Our approach is to instead decompose $X_{\log}$ as

$$X_{\log} = X_{\text{norm}} + X_{\text{corr}}, \tag{9}$$

where we assume $X_{\text{norm}} \sim \mathcal{N}(0,\sigma^2)$ and that X_{corr} is a zero-mean "correction" variable with density $C_\sigma(X)$. The two variables are added (i.e., their distributions convolve) to form $X_{\log}$. Using $X_{\text{corr}} \sim C_\sigma(X)$, the acceptance test is now

$$\Delta + X_{\log} = (\Delta + X_{\text{norm}}) + X_{\text{corr}} = \Delta^* + X_{\text{corr}} > 0. \tag{10}$$

Therefore, assuming the variance of Δ^* is small enough, if we have an estimate of Δ^* from the current data minibatch, we test acceptance by adding a random variable X_{corr} and then accept θ' if the result is positive (and reject otherwise).

If $\bar{\mathcal{N}}(0,\sigma^2(\Delta^*))$ is exactly $\mathcal{N}(0,\sigma^2(\Delta^*))$, the above test is exact, and as we show in Section 4, if there is a maximum error ϵ between the CDF of $\bar{\mathcal{N}}(0,\sigma^2(\Delta^*))$ and the CDF of $\mathcal{N}(0,\sigma^2(\Delta^*))$, then our test has an error of at most ϵ relative to the full batch version.

3.3 The Correction Distribution

Our test in Equation (10) requires knowing the distribution of X_{corr}. In Section 4, we show that the test accuracy depends on the absolute error between the CDFs of $X_{\text{norm}} + X_{\text{corr}}$ and $X_{\log}$. Consequently, we need to minimize this in our construction of X_{corr}; we reduce the problem to a least squares minimization (see [Seita *et al.*, 2017], Section 4).

4 Analysis

We now derive error bounds for our M-H test and the target distribution it generates. See [Seita *et al.*, 2017] for proofs.

4.1 Bounding the Error Of Δ^* From Normal

We use the following from Corollary 2 in [Novak, 2005]:

Lemma 2. *Let $X_1,\ldots,X_n$ be a set of zero-mean, independent, identically-distributed random variables with sample mean $\bar{X}$ and sample variance s_X^2 where:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s_X = \frac{1}{n}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{\frac{1}{2}}. \tag{11}$$

Then the t-statistic $t = \bar{X}/s_X$ has a distribution which is approximately normal, with error bounded by:

$$\sup_x |\Pr(t < x) - \Phi(x)| \leq \frac{6.4\mathbb{E}[|X|^3] + 2\mathbb{E}[|X|]}{\sqrt{n}}. \tag{12}$$

Lemma 2 demonstrates that if we know $\mathbb{E}[|X|]$ and $\mathbb{E}[|X|^3]$, we can bound the error of the normal approximation, which decays as $O(n^{-\frac{1}{2}})$. Making the change of variables $y = xs_X$, Equation (12) becomes

$$\sup_y \left|\Pr(\bar{X} < y) - \Phi\left(\frac{y}{s_X}\right)\right| \leq \frac{6.4\mathbb{E}[|X|^3] + 2\mathbb{E}[|X|]}{\sqrt{n}} \tag{13}$$

showing that the distribution of $\bar{X}$ approaches the normal distribution $\mathcal{N}(0,s_X)$ whose standard deviation is s_X, as measured from the sample. To apply this to our test, let $X_i = \Lambda_i(\theta,\theta') - \Lambda(\theta,\theta')$, so that the X_i are zero-mean, i.i.d. variables. If we only extract a subset of b samples from our minibatch, then $\bar{X} = \Delta^*(\theta,\theta') - \Delta(\theta,\theta')$, so that $s_X = s_{\Delta^*}$. We can now substitute into Equation (13):

Corollary 1.

$$\sup_y \left|\Pr(\Delta^* < y) - \Phi\left(\frac{y-\Delta}{s_{\Delta^*}}\right)\right| \leq \frac{6.4\mathbb{E}[|X|^3] + 2\mathbb{E}[|X|]}{\sqrt{b}} \tag{14}$$

where the upper bound can be expressed as $\epsilon(\theta,\theta',b)$. Corollary 1 shows that the distribution of Δ^* approximates a Normal distribution with mean Δ and variance $s^2_{\Delta^*}$. Furthermore, it bounds the error with *estimable quantities*: both $\mathbb{E}[|X|]$ and $\mathbb{E}[|X|^3]$ can be estimated as means of $|\Lambda_i - \Lambda|$ and $|\Lambda_i - \Lambda|^3$, respectively, on each minibatch.

4.2 Adding Random Variables

By applying Lemma 4 from [Seita *et al.*, 2017], we have:

Corollary 2. *If* $\sup_y |\Pr(\Delta^* < y) - \Phi(\frac{y-\Delta}{s_{\Delta^*}})| \leq \epsilon(\theta,\theta',b)$, *then*

$$\sup_y |\Pr(\Delta^* + X_{\text{nc}} + X_{\text{corr}} < y) - S(y-\Delta)| \leq \epsilon(\theta,\theta',b)$$

where $S(x)$ is the standard logistic function, and X_{nc} and X_{corr} are generated as per Algorithm 1.

Corollary 2 shows that the bounds from Section 4.1 are preserved after adding random variables, so our test remains accurate. Furthermore, bounds on the error of an M-H test imply bounds on the stationary distribution of the Markov chain under appropriate conditions [Korattikara *et al.*, 2014].

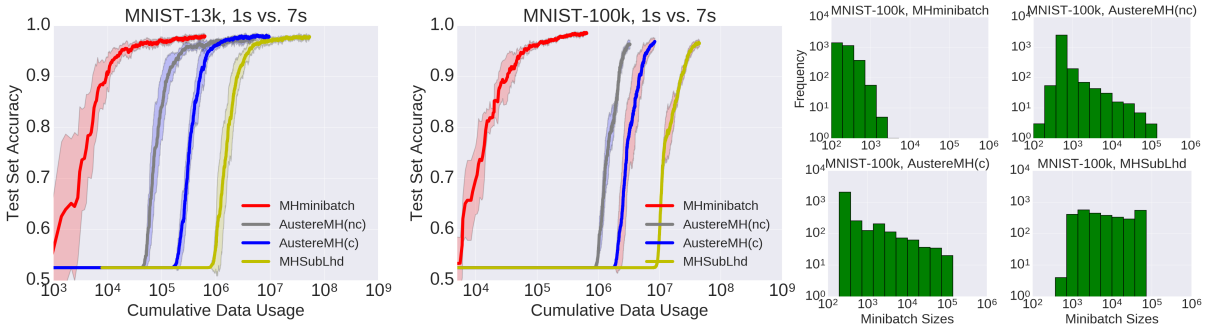

Figure 1: Classification accuracy on 1s vs 7s for MNIST-13k (left) and MNIST-100k (middle) as a function of cumulative data usage; error regions indicate one standard deviation. Right: four subplots of minibatch sizes in a representative trial for each method.

Method	Average of MB Sizes
MHMINIBATCH	182.3 ± 11.4
AUSTEREMH(C)	13540.5 ± 1521.4
MHSUBLHD	65758.9 ± 3222.6

Table 1: Average minibatch sizes ($\pm$ one standard deviation) on Gaussian mixture models, over 10 trials (3000 samples each).

Method/Data	MNIST-13k	MNIST-100k
MHMINIBATCH	125.4 ± 9.2	216.5 ± 7.9
AUSTEREMH(NC)	973.8 ± 49.8	1098.3 ± 44.9
AUSTEREMH(C)	1924.3 ± 52.4	2795.6 ± 364.0
MHSUBLHD	10783.4 ± 78.9	14977.3 ± 582.0

Table 2: Average minibatch sizes ($\pm$ one standard deviation) for LR on MNIST-13k (10 trials) and MNIST-100k (5 trials).

5 Experiments

We compare with the most similar prior works [Korattikara *et al.*, 2014] and [Bardenet *et al.*, 2014] on Mixtures of Gaussians and Logistic Regression. For full experiment details, see [Seita *et al.*, 2017]; here we provide a brief summary.

5.1 Mixture Of Gaussians

This model is adapted from [Welling and Teh, 2011] by increasing the number of samples to 1 million. The parameters are $\theta = \langle \theta_1, \theta_2 \rangle$, and the generation process is

$$\begin{aligned} \theta &\sim \mathcal{N}(0, \text{diag}(\sigma_1^2, \sigma_2^2)) \\ x_i &\sim 0.5 \cdot \mathcal{N}(\theta_1, \sigma_x^2) + 0.5 \cdot \mathcal{N}(\theta_1 + \theta_2, \sigma_x^2). \end{aligned} \tag{15}$$

We set $\sigma_1^2 = 10, \sigma_2^2 = 1$, $\sigma_x^2 = 2$, and $\theta = \langle 0, 1 \rangle$. We are interested in large-scale problems and sampled 1,000,000 points to form posterior $\propto p(\theta) \prod_{i=1}^{1,000,000} p(x_i|\theta)^{1/K}$ with the same prior from Equation (15). This produces a much sharper posterior with two very narrow peaks. To reproduce the original posterior, we set temperature $K = 10,000$.

We benchmark with AUSTEREMH(C) and MHSUBLHD. All methods collect 3000 samples using a random walk proposer with covariance matrix $\text{diag}(0.15, 0.15)$, which means the M-H test is responsible for shaping the sample distribution. Section 6.1 in [Seita *et al.*, 2017] indicates that there are no obvious differences in the posterior samples. Table 1 shows that MHMINIBATCH dominates in terms of efficiency with about a 100x improvement over alternative methods.

5.2 Logistic Regression

We test logistic regression for binary classification of 1s vs. 7s on MNIST [LeCun and Cortes, 1998] and a subset of infinite MNIST [Loosli *et al.*, 2007]. For the former, extracting all 1s and 7s resulted in 13,000 training samples ("MNIST-13k"), and for the latter, we used 87,000 additional (augmented) 1s and 7s to get 100,000 total ("MNIST-100k"). Both datasets use the same test set, with 2,163 samples. For all methods, we impose a uniform prior and again use a random walk proposer with covariance $0.05I$ for MNIST-13k and $0.01I$ for MNIST-100k. The temperature is $K = 100$.

The first two subplots of Figure 1 display the prediction accuracy on both datasets for all methods as a function of the cumulative training points processed (note that the methods consume different amounts of data). To generate the curves, for each sample θ_t, $t \in \{1, \ldots, T\}$, we use θ_t as the logistic regression parameter. The results suggest that our test obtains convergence more than 10x faster than the others.

Table 2 contains the average of the average minibatch sizes ($\pm$ one standard deviation) across all trials; see [Seita *et al.*, 2017] for detailed histograms and the right subplot of Figure 1 for a representative result. MHMINIBATCH, with sizes of 125.4 and 216.5 for MNIST-13k and MNIST-100k, respectively, consumes more than 7x and 4x fewer data points than the next-best method, AUSTEREMH(NC). We note that both AUSTEREMH(NC) and MHSUBLHD require computing $\log p(x_i|\theta)$ and $\log p(x_i|\theta')$ for all x_i each iteration.

6 Conclusions

We have derived a minibatch M-H which approximates full data tests and presented promising theoretical and empirical results. A priority is to integrate the test with HMC and variants [Homan and Gelman, 2014; Tripuraneni *et al.*, 2017]. Another idea is to merge our test with [Korattikara *et al.*, 2014] by applying both each iteration and utilizing variance reduction techniques [Chen and Ghahramani, 2016].

Acknowledgments

We thank the reviewers who provided helpful comments. Daniel Seita is supported by an NPSC fellowship.

References

[Ahn *et al.*, 2012] Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In *International Conference on Machine Learning (ICML)*, 2012.

[Bardenet *et al.*, 2014] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards Scaling up Markov chain Monte Carlo: An Adaptive Subsampling Approach. In *International Conference on Machine Learning (ICML)*, 2014.

[Bardenet *et al.*, 2017] Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov Chain Monte Carlo Methods for Tall Data. *Journal of Machine Learning Research (JMLR)*, 2017.

[Barker, 1965] A. A. Barker. Monte-Carlo Calculations of the Radial Distribution Functions for a Proton-Electron Plasma. *Australian Journal of Physics*, 18:119–133, 1965.

[Betancourt, 2017] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[Brooks *et al.*, 2011] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[Chen and Ghahramani, 2016] Yutian Chen and Zoubin Ghahramani. Scalable Discrete Sampling as a Multi-Armed Bandit Problem. In *International Conference on Machine Learning (ICML)*, 2016.

[Chen *et al.*, 2014] T. Chen, E.B. Fox, and C. Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*, 2014.

[Dupuy and Bach, 2016] Christophe Dupuy and Francis Bach. Online but Accurate Inference for Latent Variable Models with Local Gibbs Sampling. *arXiv preprint arXiv:1603.02644*, 2016.

[Gilks and Spiegelhalter, 1996] W.R. Gilks and DJ Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.

[Hastings, 1970] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:97–109, 1970.

[Homan and Gelman, 2014] Matthew D. Homan and Andrew Gelman. The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research (JMLR)*, 15(1):1593–1623, January 2014.

[Korattikara *et al.*, 2014] Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *International Conference on Machine Learning (ICML)*, 2014.

[LeCun and Cortes, 1998] Yann LeCun and Corinna Cortes. MNIST Handwritten Digit Database. 1998.

[Loosli *et al.*, 2007] Gaëlle Loosli, Stéphane Canu, and Léon Bottou. Training Invariant Support Vector Machines using Selective Sampling. In Léon Bottou, Olivier Chapelle, Dennis DeCoste, and Jason Weston, editors, *Large Scale Kernel Machines*, pages 301–320. MIT Press, Cambridge, MA., 2007.

[Maclaurin and Adams, 2014] Dougal Maclaurin and Ryan P. Adams. Firefly Monte Carlo: Exact MCMC with Subsets of Data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

[Metropolis *et al.*, 1953] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21, 1953.

[Neal, 2010] Radford M. Neal. MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

[Novak, 2005] Y. Novak. On Self-Normalized Sums and Student's Statistic. *Theory of Probability and its Applications*, 49(2):336–344, 2005.

[Roberts and Rosenthal, 2001] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, 16(4):351–367, 2001.

[Seita *et al.*, 2017] Daniel Seita, Xinlei Pan, Haoyu Chen, and John Canny. An Efficient Minibatch Acceptance Test for Metropolis Hastings. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

[Tripuraneni *et al.*, 2017] Nilesh Tripuraneni, Mark Rowland, Zoubin Ghahramani, and Richard Turner. Magnetic Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*, 2017.

[Wang and Uhlenbeck, 1945] M.C. Wang and G.E. Uhlenbeck. *On the Theory of the Brownian Motion II.* Reviews of Modern Physics, 1945.

[Welling and Teh, 2011] Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *International Conference on Machine Learning (ICML)*, 2011.