An Efficient Minibatch Acceptance Test for Metropolis-Hastings*

Daniel Seita¹, Xinlei Pan¹, Haoyu Chen¹, John Canny¹,²
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 orders-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 \( \theta \) 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 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 \( \theta_{t+1} \) at time \( t+1 \) is generated using a candidate \( \theta' \) 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:

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

where \( \wedge \) denotes \( \min(a, b) \). With probability \( \alpha(\theta_t, \theta') \), we accept \( \theta' \) and set \( \theta_{t+1} = \theta' \); otherwise \( \theta_{t+1} = \theta_t \).

We consider functions other than the classical Metropolis test and set \( \theta_{t+1} = \theta' \) 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

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

and \( \psi(u, \theta, \theta') = \log \left( u \frac{q(\theta'|\theta)q(\theta)}{q(\theta'|\theta)p(\theta)} \right) \). 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, \( p(x_i|\theta)K^{-1} \), 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 \( \{x_{i1}, \ldots, x_{ib}\} \) can be used:

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

Finally, define \( \Lambda_i(\theta, \theta') = N \log \left( \frac{p(x_i|\theta')}{p(x_i|\theta)} \right) \), 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 \( \Delta \) and \( \Delta^* \), respectively. First, \( \Delta \) is

\[
\Delta(\theta, \theta') = \frac{p_0(\theta') \prod_{i=1}^{N} p(x_i|\theta')q(\theta|\theta')}{p_0(\theta) \prod_{i=1}^{N} p(x_i|\theta)q(\theta|\theta)}
\]

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 \( \theta' \) 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 \( \theta \).

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.

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 \( g(s) = \exp(s)g(-s) \), 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 anti-symmetric) and a correction (Section 3.3). The Logistic CDF \( L \) and Normal CDF \( \Phi \) 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 \( \theta \) and \( \theta' \), and have \( V \sim U(0, 1) \). We accept \( \theta' \) 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 \( \theta' \) 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 \( \Delta \). In the limit, we can write
\[
\Delta^* = \Delta + X_{\text{norm}}, \quad X_{\text{norm}} \sim N(0, \sigma^2(\Delta^*)), \tag{8}
\]
where \( N(0, \sigma^2(\Delta^*)) \) 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_{\text{log from now on. In [Bardenet et al., 2017], it was proposed to use \( \Delta^* \) in a Barker test, and tolerate the fixed error between the logistic and normal distributions. Our approach is to instead decompose \( X_{\text{log}} \) as
\[
X_{\text{log}} = X_{\text{norm}} + X_{\text{corr}}, \tag{9}
\]
where we assume \( X_{\text{norm}} \sim N(0, \sigma^2) \) and that \( X_{\text{corr}} \) is a zero-mean “correction” variable with density \( C_{\text{n}}(X) \). The two variables are added (i.e., their distributions convolve) to form \( X_{\text{log}} \). Using \( X_{\text{corr}} \sim C_{\text{n}}(X) \), the acceptance test is now
\[
\Delta + X_{\text{log}} = (\Delta + X_{\text{norm}}) + X_{\text{corr}} = \Delta^* + X_{\text{corr}} > 0. \tag{10}
\]
Therefore, assuming the variance of \( \Delta^* \) is small enough, if we have an estimate of \( \Delta^* \) from the current data minibatch, we test acceptance by adding a random variable \( X_{\text{corr}} \) and then accept \( \theta' \) if the result is positive (and reject otherwise).

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

### 3.3 The Correction Distribution

Our test in Equation (10) requires knowing the distribution of \( X_{\text{corr}} \). 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 \( X_{\text{log}} \). Consequently, we need to minimize this in our construction of \( X_{\text{corr}} \); 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 \( \Delta^* \) 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:
\[
\sup_{y} |\Pr(t < x) - \Phi(x)| \leq \frac{6.4E[|X|^3] + 2E[|X|]}{\sqrt{n}}. \tag{12}
\]

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 \( O(n^{-\frac{3}{2}}) \). Making the change of variables \( y = xs_X \), Equation (12) becomes
\[
\sup_{y} |\Pr(\bar{X} < y) - \Phi\left(\frac{y}{s_X}\right)| \leq \frac{6.4E[|X|^3] + 2E[|X|]}{\sqrt{n}}. \tag{13}
\]
showing that the distribution of \( \bar{X} \) approaches the normal distribution \( N(0, s_X) \) 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.**
\[
\sup_{y} |\Pr(\Delta^* < y) - \Phi\left(\frac{y - \Delta}{s_{\Delta^*}}\right)| \leq \frac{6.4E[|X|^3] + 2E[|X|]}{\sqrt{b}}. \tag{14}
\]
where the upper bound can be expressed as \( \epsilon(\theta, \theta', b) \).

**Corollary 2.** If \( \sup_{y} |\Pr(\Delta^* < y) - \Phi\left(\frac{y - \Delta}{s_{\Delta^*}}\right)| \leq \epsilon(\theta, \theta', b) \), then
\[
\sup_{y} |\Pr(\Delta^* + X_{nc} + X_{corr} < y) - S(y - \Delta)| \leq \epsilon(\theta, \theta', b)
\]
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].
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 $\theta = (\theta_1, \theta_2)$, and the generation process is

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

We set $\sigma_1^2 = 10, \sigma_2^2 = 1, \sigma_x^2 = 2$, and $\theta = (0, 1)$. 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 MMINIBATCH 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 [Iosif 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 0.05$I$ for MNIST-13k and 0.01$I$ 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 $\theta_i$, $t \in \{1, \ldots, T\}$, we use $\theta_i$ 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 (± one standard deviation) for LR on MNIST-13k (10 trials) and MNIST-100k (5 trials).

<table>
<thead>
<tr>
<th>Method</th>
<th>Average of MB Sizes</th>
<th>MNIST-13k</th>
<th>MNIST-100k</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMINIBATCH</td>
<td>182.3 ± 11.4</td>
<td>125.4 ± 9.2</td>
<td>216.5 ± 7.9</td>
</tr>
<tr>
<td>AUSTEREMH(C)</td>
<td>13540.5 ± 1521.4</td>
<td>973.8 ± 49.8</td>
<td>1098.3 ± 44.9</td>
</tr>
<tr>
<td>MHSUBLHD</td>
<td>65758.9 ± 3222.6</td>
<td>1924.2 ± 63.4</td>
<td>2795.6 ± 364.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method/Data</th>
<th>MNIST-13k</th>
<th>MNIST-100k</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMINIBATCH</td>
<td>10783.4 ± 78.9</td>
<td>14977.3 ± 582.0</td>
</tr>
<tr>
<td>AUSTEREMH(NC)</td>
<td>364</td>
<td>9</td>
</tr>
<tr>
<td>AUSTEREMH(C)</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>MHSUBLHD</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

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

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


