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 yield variable data consumed per sample, with 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. The acceptance test is usually a Metropolis test [Metropolis et al., 1953, Hastings, 1970].

Many state-of-the-art machine learning methods, and deep learning in particular, are based on minibatch updates (such as SGD) to a model. Minibatch 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. Recent results from [Korattikara et al., 2014] and [Bardenet et al., 2014] perform approximate (bounded error) acceptance tests using subsets of the full dataset. The amount of data consumed for each test varies significantly from one minibatch to the next. By contrast, [Maclaurin and Adams, 2014, Bardenet et al., 2016] perform exact tests but require a lower bound on the parameter distribution across its domain. The amount of data reduction depends on the accuracy of this bound, 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], which makes our test naturally error tolerant. The idea of using a noise-tolerant Barker's test function was suggested but not explored empirically in [Bardenet et al., 2016] section 6.3. But the asymptotic test statistic CDF and the Barker function are different, which leads to fixed errors for the approach in [Bardenet et al., 2016]. 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] and [Bardenet et al., 2014] which use tail bounds and provide only approximate tests even with arbitrarily large batches of data. Our test is also exact under the assumptions of Korattikara et al. [2014] that the log probability ratios of batches are normally distributed about their mean. Rather than tail bounds, our approach uses moment estimates from the data to determine how far the minibatch posteriors deviate from a normal distribution. These bounds carry through to the

overall accuracy of the test.

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). It's not clear at first why this quantity should be bounded, but it is natural for well-specified models running Metropolis-Hastings sampling with optimal proposals [Roberts and Rosenthal, 2001] on a full dataset. If the posterior parameter distribution is a unit-variance normal distribution, then the posterior for N samples will have variance $1/N$. There is simply not enough information in $M \ll N$ samples to locate and efficiently sample from this posterior. This is not a property of any particular proposal or test, but of the information carried by the data. The variance condition succinctly captures the condition that the minibatch carries enough information to generate a sample. While we cannot expect to generate independent samples from the posterior using only a small subset of the data, there are three situations where we can exploit small minibatches:

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 $1/K^2$. As we demonstrate in Section 6.2, higher temperature can be advantageous for parameter exploration.
2. For continuous distributions, reduce the proposal step size (i.e. generate correlated samples). The variance of the log acceptance probability scales as the square of proposal step size.
3. Utilize Hamiltonian Dynamics 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 terms of the information carried by the samples, this approach is not limited by the data in a particular minibatch since momentum is carried over time and "remembered" across multiple minibatches.

We note that case two above is characteristic of Gibbs samplers applied to large datasets [Dupuy and Bach, 2016]. Such samplers represent a model posterior via counts over an entire dataset of N samples. When a minibatch of M samples is used to update the model, the counts for these samples only are updated. This creates "steps" of $O(M/N)$ in the model parameters, and correlated samples from the model posterior. Correlated samples are still very useful in high-dimensional ML problems with multi-modal posteriors since they correspond to a finer-scale random walk through the posterior landscape. The contributions of this paper are as follows:

- We develop a new, more efficient (in samples per test) minibatch acceptance test with quantifiable error bounds. The test uses a novel additive correction variable to implement a Barker test based on minibatch mean and variance.
- We compare our new test and prior approaches on several datasets. We demonstrate several order-of-magnitude improvements in sample efficiency, and that the batch size distribution is short-tailed.

2 PRELIMINARIES

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.

The acceptance test guarantees detailed balance, which means $p(\theta)p(\theta'|\theta) = p(\theta')p(\theta|\theta')$, where $p(\theta'|\theta)$ is the probability of a transition from state θ to θ'. Here, $p(\theta'|\theta) = q(\theta'|\theta)\alpha(\theta, \theta')$. This condition, together with ergodicity, guarantees that the Markov chain has a unique stationary distribution $\pi(\theta) = p(\theta)$. For Bayesian inference, the target distribution is $p(\theta|x_1, \ldots, x_N)$. The acceptance probability is now:

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

where $p_0(\theta)$ is the prior. Computing samples this way requires all N data points, but this is very expensive for large datasets.

To address this challenge, [Korattikara et al., 2014, Bardenet et al., 2014] perform approximate Metropolis-Hasting tests using sequential hypothesis testing. 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 is increased and the test repeated. This process continues until a decision. These methods either invoke the asymptotic CLT and assume that finite batch

errors are normally distributed [Korattikara et al., 2014] or use a concentration bound [Bardenet et al., 2014]. We refer to these algorithms, respectively, as AUSTEREMH and MHSUBLHD. While both show useful reductions in the number of samples required, they suffer from two drawbacks: (i) They are approximate, and always yield a decision with a finite error, (ii) They both require exact, dataset-wide bounds that depend on θ (see Section 5).[1] We discuss a worst-case scenario in Section 2.2.

2.1 NOTATION

Following [Bardenet et al., 2014], we write the test $u < \alpha(\theta, \theta')$ equivalently as $\Lambda(\theta, \theta') > \psi(u, \theta, \theta')$, where[2]

$$\begin{aligned} \Lambda(\theta, \theta') &= \sum_{i=1}^{N} \log \frac{p(x_i|\theta')}{p(x_i|\theta)}, \\ \psi(u, \theta, \theta') &= \log \left(u \frac{q(\theta'|\theta) p_0(\theta)}{q(\theta|\theta') p_0(\theta')} \right). \end{aligned} \tag{3}$$

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, it will be convenient for our analysis to define $\Lambda_i(\theta, \theta') = N \log(\frac{p(x_i|\theta')}{p(x_i|\theta)})$. Thus, $\Lambda(\theta, \theta')$ is the mean of $\Lambda_i(\theta, \theta')$ over the entire dataset, and $\Lambda^*(\theta, \theta')$ is the mean of the $\Lambda_i(\theta, \theta')$ in its minibatch.

Since minibatches contains randomly selected samples, the values Λ_i are i.i.d. random variables.[3] By the Central Limit Theorem, we expect $\Lambda^*(\theta, \theta')$ to be approximately Gaussian. The acceptance test then becomes a statistical test of the hypothesis that $\Lambda(\theta, \theta') > \psi(u, \theta, \theta')$ by establishing that $\Lambda^*(\theta, \theta')$ is substantially larger than $\psi(u, \theta, \theta')$.

2.2 A WORST-CASE GAUSSIAN EXAMPLE

Let $x_1, \ldots, x_N$ be i.i.d. $\mathcal{N}(\theta, 1)$ with known variance $\sigma^2 = 1$ and (unknown) mean $\theta = 0.5$. We use a uniform prior on θ. The log likelihood ratio is

$$\Lambda^*(\theta, \theta') = N(\theta' - \theta) \left(\frac{1}{b} \sum_{i=1}^{b} x_i^* - \theta - \frac{\theta' - \theta}{2} \right) \tag{5}$$

which is normally distributed over selection of the Normal samples x_i^*. Since the x_i^* have unit variance, their mean has variance $1/b$, and the variance of $\Lambda^*(\theta, \theta')$ is $\sigma^2(\Lambda^*) = (\theta' - \theta)^2 N^2 / b$. In order to pass a hypothesis test that $\Lambda > \psi$, there needs to be a large enough gap (several $\sigma(\Lambda^*)$) between $\Lambda^*(\theta, \theta')$ and $\psi(u, \theta, \theta')$.

The posterior is a Gaussian centered on the sample mean μ, and with variance $1/N$ (i.e., $\mathcal{N}(\mu, 1/N)$). In one dimension, an efficient proposal distribution has the same variance as the target distribution [Roberts and Rosenthal, 2001], so we use a proposal based on $\mathcal{N}(\theta, 1/N)$. It is symmetric $q(\theta'|\theta) = q(\theta|\theta')$, and since we assumed a uniform prior, $\psi(u, \theta, \theta') = \log u$. Our worst-case scenario is specified in Lemma 1.

Lemma 1. *For the model in Section 2.2, there exists a fixed (independent of N) constant c such that with probability $\geq c$ over the joint distribution of (θ, θ', u),* AUSTEREMH *and* MHSUBLHD *consume all N samples.*

Proof. See Appendix, Section A.1. □

Similar results can be shown for other distributions and proposals by identifying regions in product space (θ, θ', u) such that the hypothesis test needs to separate nearly-equal values. It follows that the accelerated tests from prior work require at least a constant fraction $\geq c$ in the amount of data consumed per test compared to full-data tests, so their speed-up is $\leq 1/c$. The issue is the use of tail bounds to separate $\Lambda - \psi$ from zero; for certain input/random u combinations, this difference can be arbitrarily close to zero. We avoid this by using the *approximately normal* variation in Λ^* to *replace* the variation due to u.

2.3 MCMC POSTERIOR INFERENCE

There is a separate line of MCMC work drawing principles from statistical physics. One can apply Hamiltonian Monte Carlo (HMC) [Neal, 2010] methods which generate high acceptance *and* distant proposals when run on full batches of data. Recently Langevin Dynamics [Welling and Teh, 2011, Ahn et al., 2012] has been applied to Bayesian estimation on minibatches of data. This simplified dynamics uses local proposals and avoids M-H tests by using small proposal steps whose acceptance approaches 1 in the limit. However, the constraint on proposal step size is severe, and the state space exploration reduces to a random walk. Full minibatch HMC for minibatches was described in [Chen

[1] We obtained the authors code for both and found that they scanned the entire dataset at each step to obtain these estimates.

[2] Our definitions differ from those in [Bardenet et al., 2014] by a factor of N to simplify our analysis later.

[3] The analysis assumes sampling with replacement although implementations on typical large datasets will approximate this by sampling without replacement.

et al., 2014] which allows momentum-augmented proposals with larger step sizes. However, step sizes are still limited by the need to run accurately without M-H tests. By providing an M-H test with similar cost to standard gradient steps, our work opens the door to applying those methods with much more aggressive step sizes without loss of accuracy.

3 A NEW MH ACCEPTANCE TEST

3.1 LOG-LIKELIHOOD RATIOS

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

$$\Delta(\theta, \theta') = \log \frac{p_0(\theta') \prod_{i=1}^N p(x_i|\theta')q(\theta|\theta')}{p_0(\theta) \prod_{i=1}^N p(x_i|\theta)q(\theta'|\theta)}, \quad (6)$$

where p_0, p, and q match the corresponding functions within Equation (2). We separate out terms dependent and independent of the data as:

$$\Delta(\theta, \theta') = \underbrace{\sum_{i=1}^N \log \frac{p(x_i|\theta')}{p(x_i|\theta)}}_{\Lambda(\theta,\theta')} - \psi(1, \theta, \theta'). \quad (7)$$

A minibatch estimator of Δ, denoted as Δ^*, is

$$\Delta^*(\theta, \theta') = \underbrace{\frac{N}{b} \sum_{i=1}^b \log \frac{p(x_i^*|\theta')}{p(x_i^*|\theta)}}_{\Lambda^*(\theta,\theta')} - \psi(1, \theta, \theta'). \quad (8)$$

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 $\Delta(\theta, \theta')$.

The key to our test is a smooth acceptance function. We consider functions other than the classical Metropolis test that satisfy the detailed balance condition needed for accurate posterior estimation. A class of suitable functions is specified as follows:

Lemma 2. *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.*

This result is used in [Barker, 1965] to define the Barker acceptance test.

3.2 BARKER (LOGISTIC) ACCEPTANCE FUNCTION

For our new MH test we use the Barker logistic [Barker, 1965] function: $g(s) = (1 + \exp(-s))^{-1}$. Straightforward arithmetic shows that it satisfies the condition in Lemma 2. It is slightly less efficient than the Metropolis test, since its acceptance rate for vanishing likelihood difference is 0.5. However we will see that its overall sample efficiency is much higher than the earlier methods. See Appendix B for additional discussion.

Assume we begin with the current sample θ and a candidate sample θ', and that $V \sim \mathcal{U}(0, 1)$ is a uniform random variable. 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 an equivalent test is to accept θ' iff

$$\Delta(\theta, \theta') > X = g^{-1}(V) \quad (9)$$

where X is a random variable with the logistic distribution (its CDF is the logistic function). To see this notice that $\frac{dV}{dX} = g'$, that g' is the density corresponding to a logistic CDF, and finally that $\frac{dV}{dX}$ is the density of X. The density of X is symmetric, so we can equivalently test whether

$$\Delta(\theta, \theta') + X > 0 \quad (10)$$

for a logistic random variable X.

3.3 A MINIBATCH ACCEPTANCE TEST

We now describe acceptance testing using the minibatch estimator $\Delta^*(\theta, \theta')$. From Equation (8), $\Delta^*(\theta, \theta')$ 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 a normal distribution by the Central Limit Theorem. We now describe this using an asymptotic argument and defer specific bounds between the CDFs of $\Delta^*(\theta, \theta')$ and a Gaussian to Section 5.

In the limit, since Δ^* is normally distributed about its mean Δ, we can write

$$\Delta^* = \Delta + X_{\text{norm}}, \quad X_{\text{norm}} \sim \bar{\mathcal{N}}(0, \sigma^2(\Delta^*)), \quad (11)$$

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 (10) we want $\Delta + X$ for a logistic random variable X (call it $X_{\log}$ from now on). In [Bardenet et al., 2016] 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

$$X_{\log} = X_{\text{norm}} + X_{\text{corr}}, \quad (12)$$

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 $X_{\log}$. This decomposition requires an appropriate C_σ, which we derive in Section 4. Using X_{corr} samples from $C_\sigma(X)$, the acceptance test is now

$$\Delta + X_{\log} = (\Delta + X_{\text{norm}}) + X_{\text{corr}} = \Delta^* + X_{\text{corr}} > 0. \tag{13}$$

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 5, 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.

4 THE CORRECTION DISTRIBUTION

Our test in Equation (13) requires knowing the distribution of X_{corr}. In Section 5, 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}. More formally, let $\Phi_{s_X} = \Phi(X/s_X)$ where Φ is the standard normal CDF[4], $S(X)$ be the logistic function, and $C_\sigma(X)$ be the *density* of the correction X_{corr} distribution. Our goal is to solve:

$$C^*_\sigma = \arg\min_{C_\sigma} |\Phi_\sigma * C_\sigma - S| \tag{14}$$

where $*$ denotes convolution. To compute C_σ, we assume the input Y and another variable X lie in the intervals $[-V, V]$ and $[-2V, 2V]$, respectively. We discretize the convolution by discretizing X and Y into $4N+1$ and $2N+1$ values respectively. If $i \in \{-2N, \ldots, 2N\} = \mathcal{I}$ and $j \in \{-N, \ldots, N\} = \mathcal{J}$, then we can write $X_i = i(V/N)$ and $Y_j = j(V/N)$, and the objective can be written as:

$$C^*_\sigma = \arg\min_{C_\sigma} \max_{i \in \mathcal{I}} \left| \sum_{j \in \mathcal{J}} \Phi_\sigma(X_i - Y_j) C_\sigma(Y_j) - S(X_i) \right|.$$

Now define matrix M and vectors u and v such that $M_{ij} = \Phi_\sigma(X_i - Y_j)$, $u_j = C_\sigma(Y_j)$, and $v_i = S(X_i)$, where the indices i and j are appropriately translated to be non-negative for M, u, and v. The problem is now to minimize $\|Mu - v\|_\infty$ with the density non-negative constraint $u > 0$. We approximate this with least squares:

$$u^* = \arg\min_u \ \|Mu - v\|_2^2 + \lambda \|u\|_2^2, \tag{15}$$

[4]Hence, Φ_{s_X} is the CDF of a zero-mean Gaussian with standard deviation s_X.

Algorithm 1 MHMINIBATCH acceptance test.

Input: number of samples T, minibatch size m, error bound δ, pre-computed correction $C_1(X)$ distribution, initial sample θ_1.
Output: a chain of T samples $\{\theta_1, \ldots, \theta_T\}$.
for $t = 1$ **to** T **do**
  -Propose a candidate θ' from proposal $q(\theta'|\theta_t)$.
  -Draw a minibatch of m points $\{x_1^*, \ldots, x_m^*\}$.
  -Compute $\Delta^*(\theta_t, \theta')$ and sample variance $s^2_{\Delta^*}$.
  -Estimate moments $\mathbb{E}[|\Lambda_i - \Lambda|]$ and $\mathbb{E}[|\Lambda_i - \Lambda|^3]$ from the sample, and error ϵ from Corollary 1.
  while $s^2_{\Delta^*} \geq 1$ **or** $\epsilon > \delta$ **do**
    -Draw m more samples to augment the minibatch, update Δ^*, $s^2_{\Delta^*}$ and ϵ estimates.
  end while
  -Draw $X_{\text{nc}} \sim \mathcal{N}(0, 1 - s^2_{\Delta^*})$ and $X_{\text{corr}} \sim C_1(X)$.
  if $\Delta^* + X_{\text{nc}} + X_{\text{corr}} > 0$ **then**
    -Accept the candidate, $\theta_{t+1} = \theta'$.
  else
    -Reject and re-use the old sample, $\theta_{t+1} = \theta_t$.
  end if
end for

with regularization λ. The solution is well-known from the normal equations ($u^* = (M^T M + \lambda I)^{-1} M^T v$) and in practice yields an acceptable L_∞ norm.

With this approach, there is no guarantee that $u^* \geq 0$. However, we have some flexibility in the choice of σ in Equation (14). As we decrease the variance of X_{norm}, the variance of X_{corr} grows by the same amount and is in fact the result of convolution with a Gaussian whose variance is the difference. Thus as σ decreases, $C_\sigma(X)$ grows and approaches the derivative of a logistic function at $\sigma = 0$. It retains some weak negative values for $\sigma > 0$ but removal of those leads to small error. We use $N = 4000$ and $\lambda = 10$ for our experiments, which empirically provided excellent performance. See Table 3 in Appendix C.1 for detailed L_∞ errors for different settings. Algorithm 1 describes our procedure, MHMINIBATCH. A few points:

- It uses an adaptive step size so as to use the smallest possible average minibatch size. Unlike previous work, the size distribution is short-tailed.
- An additional normal variable X_{nc} is added to Δ^* to produce a variable with unit variance. This is not mathematically necessary, but allows us to use a single correction distribution C_1 with $\sigma = 1$ for X_{corr}, saving on memory footprint.
- The sample variance of Δ^* is denoted as $s^2_{\Delta^*}$ and is proportional to $\|\theta' - \theta\|_2^2$.

5 ANALYSIS

We now derive error bounds for our M-H test and the target distribution it generates. In Section 5.1, we present bounds on the absolute and relative error (in terms of the CDFs) of the distribution of Δ^* versus a Gaussian. We then show in Section 5.2 that these bounds are preserved after the addition of other random variables (e.g., X_{nc} and X_{corr}). It then follows that the acceptance test has the same error bound.

5.1 BOUNDING THE ERROR OF Δ^* FROM A GAUSSIAN

We use the following quantitative central-limit result:

Lemma 3. *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:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s_X = \frac{1}{n}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{\frac{1}{2}}. \quad (16)$$

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

$$\sup_x |\Pr(t < x) - \Phi(x)| \leq \frac{6.4\mathbb{E}[|X|^3] + 2\mathbb{E}[|X|]}{\sqrt{n}}. \quad (17)$$

Proof. See Appendix, Section A.2. □

Lemma 3 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 (17) becomes

$$\sup_y \left|\Pr(\bar{X} < y) - \Phi\left(\frac{y}{s_X}\right)\right| \leq \frac{6.4\mathbb{E}[|X|^3] + 2\mathbb{E}[|X|]}{\sqrt{n}} \quad (18)$$

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 instead of all n samples, we only extract a subset of b samples corresponding to our minibatch, we can connect $\bar{X}$ with our Δ^* term: $\bar{X} = \Delta^*(\theta, \theta') - \Delta(\theta, \theta')$, so that $s_X = s_{\Delta^*}$. We can now substitute into Equation (18) and displace by the mean, giving:

Corollary 1.

$$\sup_y \left|\Pr(\Delta^* < y) - \Phi\left(\frac{y - \Delta}{s_{\Delta^*}}\right)\right| \leq \frac{6.4\mathbb{E}[|X|^3] + 2\mathbb{E}[|X|]}{\sqrt{b}} \quad (19)$$

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. We expect this will often be accurate enough on minibatches with hundreds of points, but otherwise bootstrap CIs can be computed.

5.2 ADDING RANDOM VARIABLES

We next relate the CDFs of distributions and show that bounds are preserved after adding random variables.

Lemma 4. *Let $P(x)$ and $Q(x)$ be two CDFs satisfying $\sup_x |P(x) - Q(x)| \leq \epsilon$ with x in some real range. Let $R(y)$ be the* density *of another random variable y. Let P' be the convolution $P * R$ and Q' be the convolution $Q * R$. Then $P'(z)$ (resp. $Q'(z)$) is the CDF of sum $z = x + y$ of independent random variables x with CDF $P(x)$ (resp. $Q(x)$) and y with density $R(y)$. Then*

$$\sup_x |P'(x) - Q'(x)| \leq \epsilon. \quad (20)$$

Proof. See Appendix, Section A.3. □

From Lemma 4, we have the following Corollary:

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.

Proof. See Appendix, Section A.4. □

Corollary 2 shows that the bounds from Section 5.1 are preserved after adding random variables, so our test remains accurate. In fact we can do better ($O(n^{-1})$ instead of $O(n^{-1/2})$) by using a more precise limit distribution under an additional assumption. We review this in Appendix A.5.

5.3 BOUNDS ON THE STATIONARY DISTRIBUTION

Bounds on the error of an M-H test imply bounds on the stationary distribution of the Markov chain under appropriate conditions. Such bounds were derived in both [Korattikara et al., 2014] and [Bardenet et al., 2014]. We include the result from [Korattikara et al., 2014] (Theorem 1) here: Let $d_v(P, Q)$ denote the total variation distance

between two distributions P and Q. Let $\mathcal{T}_0$ denote the transition kernel of the exact Markov chain, $\mathcal{S}_0$ denote the exact posterior distribution, and $\mathcal{S}_\epsilon$ denote the stationary distribution of the approximate transition kernel.

Lemma 5. *If $\mathcal{T}_0$ satisfies the contraction condition $d_v(P\mathcal{T}_0, \mathcal{S}_0) < \eta d_v(P, \mathcal{S}_0)$ for some constant $\eta \in [0, 1)$ and all probability distributions P, then*

$$d_v(S_0, S_\epsilon) \leq \frac{\epsilon}{1-\eta} \tag{21}$$

where ϵ is the bound on the error in the acceptance test.

6 EXPERIMENTS

Here we compare with the most similar prior works [Korattikara et al., 2014] and Bardenet et al. [2014]. In [Korattikara et al., 2014], an asymptotic CLT is used to argue that a modified standard M-H test can be used on subsets of the data. This assumes knowledge of dataset-wide mean μ_{std} each iteration (it depends on θ). Determining μ_{std} exactly requires a scan over the entire dataset, or some model-specific bounds. [Korattikara et al., 2014] also propose a conservative variant which assumes $\mu_{\text{std}} = 0$ and avoids the scan. We refer to the conservative version as AUSTEREMH(C) and the non-conservative variant as AUSTEREMH(NC). We analyze both in this section.

In [Bardenet et al., 2014] concentration bounds are used with a similar modification to the standard M-H test (MHSUBLHD method). For MHSUBLHD, the required global bound is denoted $C_{\theta,\theta'}$ which once again depends on θ and so must be recomputed at each step, or estimated in a model-specific way. We obtained sample code for both methods from the authors, and found that both AUSTEREMN(NC) and MHSUBLHD scanned the entire dataset at each iteration to derive these bounds. We do not include the cost of doing this in our experiments, since otherwise there would be no improvement over testing the full dataset. However, it should be kept in mind that such bounds must be provided to these methods. Our test by contrast uses a quantitative form of the CLT which rely on measurable statistics from a *single* minibatch. It therefore requires no dataset-wide scans, and can be used, e.g. on streams of data.

In Sections 6.1 and 6.2, we benchmark MHMINIBATCH against MHSUBLHD, AUSTEREMH(C) and AUSTEREMH(NC). Hyperparameters for the latter were optimized using a grid-search over minibatch sizes m and per-test thresholds ϵ described in Appendix C.2.1. Throughout our descriptions, we refer to a *trial* as the period when an algorithm collects all its desired samples $\{\theta_1, \ldots, \theta_T\}$, generally with $T = 3000$ or $T = 5000$.

6.1 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

$$\begin{aligned} \theta &\sim \mathcal{N}(0, \text{diag}(\sigma_1^2, \sigma_2^2)) \\ x_i &\sim 0.5 \cdot \mathcal{N}(\theta_1, \sigma_x^2) + 0.5 \cdot \mathcal{N}(\theta_1 + \theta_2, \sigma_x^2). \end{aligned} \tag{22}$$

We set $\sigma_1^2 = 10, \sigma_2^2 = 1$ and $\sigma_x^2 = 2$. We fix $\theta = \langle 0, 1 \rangle$. The original paper sampled 100 data points and estimated the posterior. We are interested in performance on larger problems and so sampled 1,000,000 points to form the posterior of $p(\theta) \prod_{i=1}^{1,000,000} p(x_i|\theta)^{1/K}$ with the same prior from Equation (22). This produces a much sharper posterior with two very narrow peaks. Our goal is to reproduce the original posterior, so we adjust the temperature to $K = 10,000$. Taking logs, we get the target as shown in the far right of Figure 1.

We benchmark with AUSTEREMH(C) and MHSUBLHD. We initialized MHMINIBATCH and MHSUBLHD with $m = 50$. For AUSTEREMH(C), we set the error bound ϵ to 0.005. For MHSUBLHD, we increase sizes geometrically with $\gamma = 1.5$ and use parameters $p = 2, \delta = 0.01$. 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.

Figure 1 shows scatter plots of the resulting θ samples for the three methods, with darker regions indicating a greater density of points. There are no obvious differences, showing that MHMINIBATCH reaches an acceptable posterior. We further measure the similarity between each set of samples and the actual posterior. Due to space constraints, results are in Appendix C.2.2.

Figure 2 shows that MHMINIBATCH dominates in terms of speed and efficiency. The histograms of the (final) minibatch sizes used each iteration show that our method consumes significantly less data; the distribution is short-tailed and the mean is 172, more than an order of magnitude better compared to the other two methods (averages are 12562 and 67508). We further ran 10 runs of mixture of Gaussians experiments and report minibatch sizes in Table 1. Sizes correspond to the running times of the methods, excluding the likelihood computation of all data points for AUSTEREMH(NC) and MHSUBLHD, which would drastically increase running time.

6.2 LOGISTIC REGRESSION

We next test logistic regression for the binary classification of 1s versus 7s on the MNIST [LeCun and Cortes,

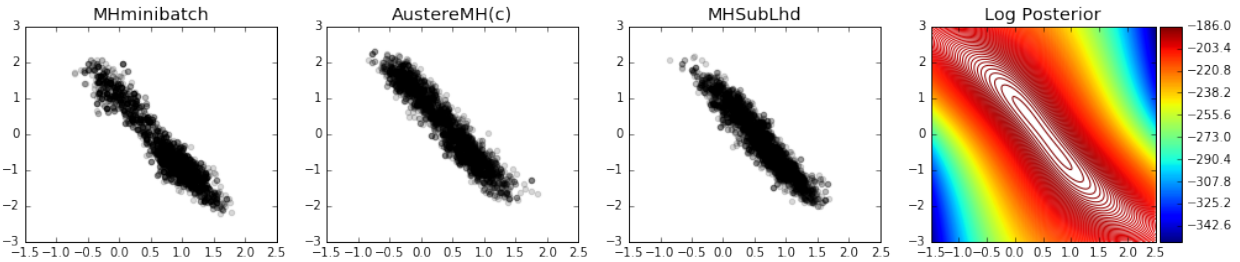

Figure 1: The log posterior contours and scatter plots of sampled θ values using different methods.

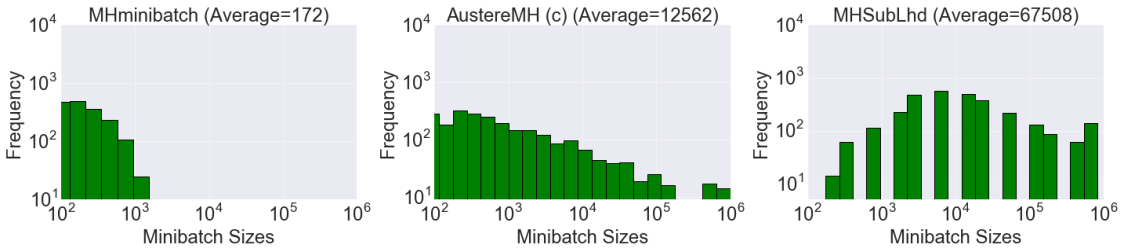

Figure 2: Minibatch sizes used in Section 6.1's experiment. The axes have the same (log-log scale) range.

Table 1: Average minibatch sizes ($\pm$ one standard deviation) on the Gaussian mixture model. The averages are taken over 10 independent trials (3000 samples each).

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

1998] dataset and (a subset of) infinite MNIST [Loosli et al., 2007]. For the former, extracting all 1s and 7s resulted in 13,000 training samples, and for the latter, we used 87,000 additional (augmented) 1s and 7s to get 100,000 training samples. Both datasets use the same test set, with 2,163 samples. Henceforth, we call them MNIST-13k and MNIST-100k, respectively.

For all methods, we impose a uniform prior on θ and again use a random walk proposer, with covariance matrix $0.05I$ for MNIST-13k and $0.01I$ for MNIST-100k. The default temperature setting is a constant at $K = 100$ for MNIST-13k and MNIST-100k. Performance of all methods implicitly relies on the step size and temperature. Setting temperature too low or step size too high will result in slow convergence for all methods. For MNIST-13k, each method generated 5000 samples for ten independent trials; due to MNIST-100k's higher computational requirements, the methods generated 3000 samples for five independent trials. For additional parameter settings and an investigation on tuning step sizes, see Appendix C.3.

For MHSUBLHD, we tried to use the provided symbolic bound for $C_{\theta,\theta'}$ described in [Bardenet et al., 2014], but it was too high and provided no performance benefit. Instead we use the empirical $C_{\theta,\theta'}$ from the entire dataset.

The first two subplots of Figure 3 display the prediction accuracy on both datasets for all methods as a function of the cumulative training points processed.[5] To generate the curves, for each of the sampled vectors θ_t, $t \in \{1, \ldots, T\}$, we use θ_t as the logistic regression parameter. The results indicate that our test is more efficient, obtaining convergence more than an order of magnitude faster than AUSTEREMH(NC) and several orders of magnitude compared to AUSTEREMH(C) and MHSUBLHD. We also observe the advantage of having higher temperature from the third plot in Figure 3, which plots average performance and one standard deviation for MHMINIBATCH over 10 trials. During the exploration period, the accuracy rapidly increases, and then after 400 samples, we switch the temperature to 1, but this requires the step size to decrease, hence the smaller changes in accuracy.

Figure 4 shows log-log histograms of minibatch sizes for the methods on MNIST-100k. (Figure 5 in Appendix C.3 contains results for MNIST-13k.) The histograms only represent one representative trial; Table 2 contains the

[5]The curves do not span the same length over the x-axis since the methods consume different amounts of data.

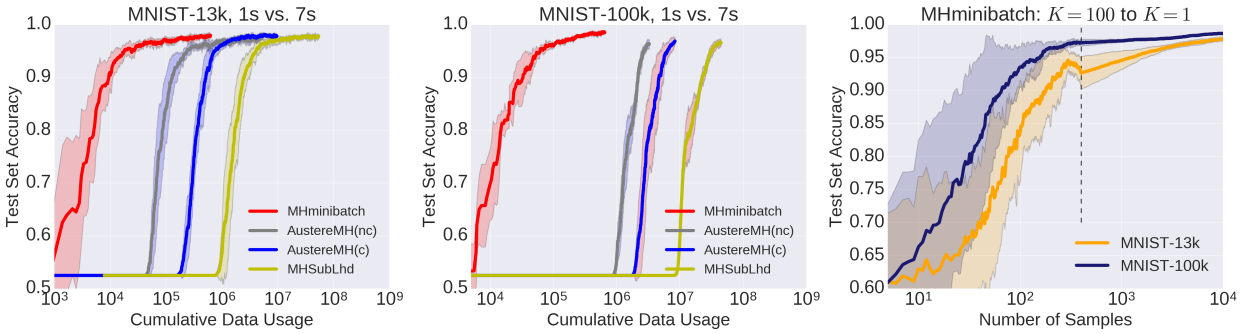

Figure 3: Binary classification accuracy of the MCMC methods on the 1s vs 7s logistic regression task for MNIST-13k (left plot) and MNIST-100k (middle plot) as a function of cumulative data usage. The right plot reports performance of MHMINIBATCH on both datasets when the temperature starts at 100 and drops to 1 after a "burn-in" period of 400 samples (vertical dashed line) of θ. For all three plots, one standard deviation is indicated by the shaded error regions.

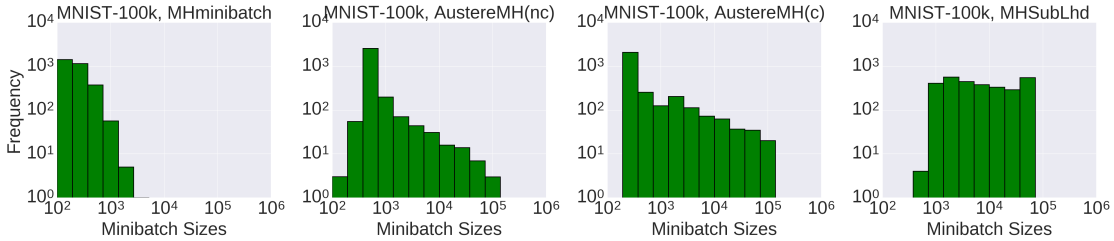

Figure 4: Minibatch sizes for a representative trial of logistic regression on MNIST-100k (analogous to Figure 2). Both axes are on a log scale and have the same ranges across the three histograms. See Section 6.2 for details.

Table 2: Average minibatch sizes ($\pm$ one standard deviation) on logistic regression on MNIST-13k and MNIST-100k. The averages are taken over 10 independent trials (5000 samples each) for MNIST-13k and 5 independent trials (3000 samples each) for MNIST-100k.

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

average of the average minibatch sizes ($\pm$ one standard deviation) across all trials. MHMINIBATCH, with average minibatch 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 reiterate, however, 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. Our results here do not count that extra data consumption. Only our method and AUSTEREMH(C) rely solely on the minibatch of data each iteration.

7 CONCLUSIONS AND DISCUSSIONS

We have derived an M-H test for minibatch MCMC which approximates full data tests. We present theoretical results and experimentally show the benefits of our test on Gaussian mixtures and a logistic regression experiment.

A priority is to extend our work to methods such as Hamiltonian Monte Carlo and Langevin Dynamics which use efficient but asymmetric proposals. While there are various approaches to symmetrizing these proposals, they have high cost in the context of minibatch MCMC. Instead we plan to extend our method to log proposal ratios which have similar structure (whole-dataset mean plus additive noise) to the log probability ratio. These can be similarly absorbed in the Barker test.

Other possibilities for future work include integrating our algorithm with [Korattikara et al., 2014] by applying both tests each iteration, utilizing the variance reduction techniques suggested in [Chen and Ghahramani, 2016], and providing recipe for how to use our algorithm following the framework of [Ma et al., 2015].

References

Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards Scaling up Markov chain Monte Carlo: An Adaptive Subsampling Approach. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov Chain Monte Carlo Methods for Tall Data. *Journal of Machine Learning Research (JMLR)*, 2016.

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

V. Bentkus, F. Gotze, and W.R.vanZwet. An Edgeworth Expansion for Symmetric Statistics. *Annals of Statistics*, 25(2), 1997.

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

T. Chen, E.B. Fox, and C. Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Yutian Chen and Zoubin Ghahramani. Scalable Discrete Sampling as a Multi-Armed Bandit Problem. In *Proceedings of the 33nd International Conference on Machine Learning, ICML*, 2016.

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

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

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

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Yann LeCun and Corinna Cortes. MNIST Handwritten Digit Database. 1998. URL `http://yann.lecun.com/exdb/mnist/`.

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.

Y. Ma, T. Chen, and E.B. Fox. A Complete Recipe for Stochastic Gradient MCMC. In *Advances in Neural Information Processing Systems 28*, 2015.

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

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.

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

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

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, 16(4):351367, 2001.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.