Appendix

This appendix is divided into three major sections. Appendix A provides the proofs that we omitted from the main text due to space constraints. Appendix B elaborates on our choice of the Barker logistic function. Finally, Appendix C presents further details on the correction distribution numerical derivation and on our three main experiments to assist understanding and reproducibility.

A PROOFS OF LEMMAS AND COROLLARIES

A.1 PROOF OF LEMMA 1

Choose \((\theta' - \theta) \in \pm \frac{1}{\sqrt{N}}[0.5, 1]\) (event 1) and \((\theta - 0.5) \in \pm \frac{1}{\sqrt{N}}[0.5, 1]\) filtered for matching sign (event 2). As discussed in Lemma 1, both \(q(\theta'|\theta)\) and \(p(\theta|x_1, \ldots, x_N)\) have variance \(1/N\). If we denote \(\Phi\) as the CDF of the standard normal distribution, then the former event occurs with probability \(p_0 = 2(\Phi(\sqrt{N}\frac{1}{\sqrt{N}}) - \Phi(\sqrt{N}\frac{0.5}{\sqrt{N}})) = 2(\Phi(1) - \Phi(0.5)) \approx 0.2997\). The latter event, because we restrict signs, occurs with probability \(p_1 = \Phi(1) - \Phi(0.5) \approx 0.14988\).

These events together guarantee that \(\Lambda^*(\theta, \theta')\) is negative by inspection of Equation (23) below. This implies that we can find a \(u \in (0, 1)\) so that \(\psi(u, \theta, \theta') = \log u < 0\) equals \(E[\Lambda^*(\theta, \theta')]\). Specifically, choose \(u_0\) to satisfy \(\log u_0 = E[\Lambda^*(\theta, \theta')]\). Using \(E[x_i^*] = 0.5\) and Equation (5), we see that

\[
\log u_0 = N(\theta' - \theta) \frac{1}{b} \cdot \mathbb{E} \left[ \sum_{i=1}^{b} x_i^* - \theta - \frac{\theta' - \theta}{2} \right] = -N(\theta' - \theta) \left( \theta - 0.5 + \frac{\theta' - \theta}{2} \right). \tag{23}
\]

Next, consider the minibatch acceptance test \(\Lambda^*(\theta, \theta') \not\approx \psi(u, \theta, \theta')\) used in Korattikara et al. [2014] and Bardenet et al. [2014], where \(\not\approx\) means “significantly different from” under the distribution over samples. This is

\[
\Lambda^*(\theta, \theta') \not\approx \psi(u_0, \theta, \theta') \iff N(\theta' - \theta) \cdot \frac{1}{b} \sum_{i=1}^{b} x_i^* - \theta - \frac{\theta' - \theta}{2} \not\approx \log u_0 \tag{24}
\]

\[
\iff \frac{1}{b} \sum_{i=1}^{b} x_i^* - \left( \theta + \frac{\theta' - \theta}{2} + \frac{\log u_0}{N(\theta' - \theta)} \right) \not\approx 0 \tag{25}
\]

\[
\iff \frac{1}{b} \sum_{i=1}^{b} x_i^* - 0.5 \not\approx 0. \tag{26}
\]

Since the \(x_i^*\) have mean 0.5, the resulting test with our chosen \(u_0\) will never correctly succeed and must use all \(N\) data points. Furthermore, if we sample values of \(u\) near enough to \(u_0\), the terms in parenthesis will not be sufficiently different from 0.5 to allow the test to succeed.

The choices above for \(\theta\) and \(\theta'\) guarantee that

\[
\log u_0 \in [-0.5, 1][0.75, 1.5] = [-1.5, -0.375]. \tag{27}
\]

Next, consider the range of \(u\) values near \(u_0\):

\[
\log u \in \log u_0 + [-0.5, 0.375]. \tag{28}
\]

The size of the range in \(u\) is at least \(\exp([-2, -1.125]) \approx [0.13534, 0.32465]\) and occurs with probability at least \(p_2 = 0.18932\). With \(u\) in this range, we rewrite the test as:

\[
\frac{1}{b} \sum_{i=1}^{b} x_i^* - 0.5 \not\approx \frac{\log u/u_0}{N(\theta' - \theta)}. \tag{29}
\]
so that, as in Equation \[26\], the LHS has expected value zero. Given our choice of intervals for the variables, we can compute the range for the right hand side (RHS) assuming\footnote{If $\theta' - \theta < 0$, then the range would be $\frac{1}{\sqrt{N}}[-0.75, 1]$ but this does not matter for the purposes of our analysis.} that $\theta' - \theta > 0$:

$$
\min\{\text{RHS}\} = \frac{-0.5}{\sqrt{N} \cdot 0.5} = -\frac{1}{\sqrt{N}} \quad \text{and} \quad \max\{\text{RHS}\} = \frac{0.375}{\sqrt{N} \cdot 0.5} = 0.75 \frac{1}{\sqrt{N}}
$$

Thus, the RHS is in $\frac{1}{\sqrt{N}}[-1, 0.75]$. The standard deviation of the LHS given the interval constraints is at least $0.5/\sqrt{b}$. Consequently, the gap between the LHS and RHS in Equation (29) is at most $2\sqrt{b/N}$ standard deviations, limiting the range in which the test will be able to “succeed” without requiring more samples.

The samples $\theta, \theta'$ and $u$ are drawn independently and so the probability of the conjunction of these events is $c = pop1p2 = 0.0085$.

### A.2 PROOF OF LEMMA 3

The following bound is given immediately after Corollary 2 from \cite{Novak2005}:

$$
-6.4E[|X|^3] - 2E[|X|] \leq \sup_x [Pr(t < x) - \Phi(x)]/\sqrt{n} \leq 1.36E[|X|^3]. \tag{31}
$$

This bound applies to $x \geq 0$. Applying the bound to $-x$ when $x < 0$ and combining with $x > 0$, we obtain the weaker but unqualified bound in Equation \[17\].

### A.3 PROOF OF LEMMA 4

We first observe that

$$
P'(z) - Q'(z) = \int_{-\infty}^{+\infty} (P(z - x) - Q(z - x))R(x)dx, \tag{32}
$$

and since $\sup_x |P(x) - Q(x)| \leq \epsilon$ it follows that $\forall z:

$$
-\epsilon = \int_{-\infty}^{+\infty} -\epsilon R(x)dx \leq \int_{-\infty}^{+\infty} (P(z - x) - Q(z - x))R(x)dx \leq \int_{-\infty}^{+\infty} \epsilon R(x)dx = \epsilon,
$$

as desired.

### A.4 PROOF OF COROLLARY 2

We apply Lemma \[4\] twice. First take:

$$
P(y) = Pr(\Delta^* < y) \quad \text{and} \quad Q(y) = \Phi \left( \frac{y - \Delta}{s_{\Delta^*}} \right)
$$

and convolve with the distribution of $X_n$ which has density $\phi(X/\sigma_n)$ where $\sigma_n^2 = 1 - s_{\Delta^*}^2$. This yields the next iteration of $P$ and $Q$:

$$
P'(y) = Pr(\Delta^* + X_{nc} < y) \quad \text{and} \quad Q'(y) = \Phi (y - \Delta)
$$

Now we convolve with the distribution of $X_{corr}$:

$$
P''(y) = Pr(\Delta^* + X_{nc} + X_{corr} < y) \quad \text{and} \quad Q''(y) = S (y - \Delta)
$$

Both steps preserve the error bound $\epsilon(\theta, \theta', b)$. Finally $S(y - \Delta)$ is a logistic CDF centered at $\Delta$, and so $S(y - \Delta) = Pr(\Delta + X_{log} < y)$ for a logistic random $X_{log}$. We conclude that the probability of acceptance for the actual test $Pr(\Delta^* + X_{nc} + X_{corr} > 0)$ differs from the exact test $Pr(\Delta + X_{log} > 0)$ by at most $\epsilon$.\footnote{If $\theta' - \theta < 0$, then the range would be $\frac{1}{\sqrt{N}}[-0.75, 1]$ but this does not matter for the purposes of our analysis.}
A.5 IMPROVED ERROR BOUNDS BASED ON SKEW ESTIMATION

We show that the CLT error bound can be improved to $O(n^{-1})$ using a more precise limit distribution under an additional assumption. Let $\mu_i$ denote the $i^{th}$ moment, and $b_i$ denote the $i^{th}$ absolute moment of $X$. If Cramer’s condition holds:

$$\lim_{t \to \infty} \sup_{i} |E[\exp(itX)]| < 1,$$

then Equation 2.2 in Bentkus et al.’s work on Edgeworth expansions [Bentkus et al., 1997] provides:

**Lemma 6.** Let $X_1, \ldots, X_n$ be a set of zero-mean, independent, identically-distributed random variables with sample mean $\bar{X}$ and with $t$ defined as in Lemma 3. If $X$ satisfies Cramer's condition, then

$$\sup_x \left| \Pr(t < x) - G(x, \frac{\mu_3}{b_2^3/2}) \right| \leq \frac{c(\epsilon, b_2, b_3, b_4, b_4+\epsilon)}{n},$$

where

$$G_n(x, y) = \Phi(x) + \frac{y(2x^2 + 1)}{6\sqrt{n}} \Phi'(x).$$

Lemma 6 shows that the average of the $X_i$ has a more precise, skewed CDF limit $G_n(x, y)$ where the skew term has weight proportional to a certain measure of skew derived from the moments: $\mu_3/b_2^{3/2}$. Note that if the $X_i$ are symmetric, the weight of the correction term is zero, and the CDF of the average of the $X_i$ converges to $\Phi(x)$ at a rate of $O(n^{-1})$.

Here the limit $G_n(x, y)$ is a normal CDF plus a correction term that decays as $n^{-1/2}$. Importantly, since $\phi''(x) = x^2\phi(x) - \phi(x)$ where $\phi(x) = \Phi'(x)$, the correction term can be rewritten giving:

$$G_n(x, y) = \Phi(x) + \frac{y}{6\sqrt{n}}(2\phi''(x) + 3\phi(x))$$

(38)

From which we see that $G_n(x, y)$ is a linear combination of $\Phi(x), \phi(x)$ and $\phi''(x)$. In Algorithm 1, we correct for the difference in $\sigma$ between $\Delta^*$ and the variance needed by $X_{\text{corr}}$ using $X_{\text{nc}}$. This same method works when we wish to estimate the error in $\Delta^*$ vs $G_n(x, y)$. Since all of the component functions of $G_n(x, y)$ are derivatives of a (unit variance) $\Phi(x)$, adding a normal variable with variance $\sigma^*$ increases the variance of all three functions to $1 + \sigma^*$. Thus we add $X_{\text{nc}}$ as per Algorithm 1 preserving the limit in Equation (38).

The deconvolution approach can be used to construct a correction variable $X_{\text{corr}}$ between $G_n(x, y)$ and $S(x)$ the standard logistic function. An additional complexity is that $G_n(x, y)$ has additional parameters $y$ and $n$. Since these act as a single multiplier $\frac{y}{\sqrt{n}}$ in Equation (38), its enough to consider a function $g(x, y')$ parametrized by $y' = \frac{y}{\sqrt{n}}$. This function can be computed and saved offline. As we have shown earlier, errors in the “limit” function propagate directly through as errors in the acceptance test. To achieve a test error of $10^{-6}$ (close to single floating point precision), we need a $y'$ spacing of $10^{-6}$. It should not be necessary to tabulate values all the way to $y' = 1$, since $y'$ is scaled inversely by the square root of minibatch size. Assuming a max $y'$ of 0.1 requires us to tabulate about 100,000. Since our $x$ resolution is 10,000, this leads to a table with about 1 billion values, which can comfortably be stored in memory. However, if $g(x, y)$ is moderately smooth in $y$, it should be possible to achieve similar accuracy with a much smaller table. We leave further analysis and experiments with $g(x, y)$ as future work.

B WHY THE BARKER LOGISTIC FUNCTION?

Regarding our choice of the Logistic function, a test function $f(x)$ for Metropolis-Hastings must satisfy Lemma 2

In addition, it must be monotone, bounded by $[0, 1]$ and be such that $\lim_{x \to -\infty} f(x) = 0$ and $\lim_{x \to \infty} f(x) = 1$. While many functions satisfy this, including the classical test $f(x) = \min\{\exp(x), 1\}$, the Logistic function is the unique function in this class which is anti-symmetric about 0.5, so it represents the (unique) CDF of a symmetric random variable. Our method requires approximating this with the sum of a Gaussian random variable (which is symmetric) and a correction. The Logistic CDF $L$ and Gaussian CDF $\Phi$ are extremely close even without correction; more precisely, the CDF error from the closest Gaussian CDF — which we numerically determined to have standard
deviation approximately 1.7 — satisfies $\sup_x |L(x) - \Phi(x/1.7)| < 0.01$. Said another way, the error between the Logistic and Gaussian CDFs is less than 1%. With our correction we can make this error orders of magnitude smaller.

While not a proof of optimality, it is unlikely that a non-symmetric test function $f(x)$ — representing a skewed variable — would do better. It would require a highly-skewed correction variable, and likely require a much narrower normal distribution (and hence more samples).

### C ADDITIONAL EXPERIMENT DETAILS

#### C.1 OBTAINING THE CORRECTION DISTRIBUTION (SECTION 4)

In Section 4 we described our derivation of the correction distribution $C_\sigma$ for random variable $X_{\text{corr}}$. Table 3 shows our $L_\infty$ error results for the convolution (Equation 14) based on various hyperparameter choices. We test using $N = 2000$ and $N = 4000$ points for discretization within a range of $X_{\text{corr}} \in [-20, 20]$, covering essentially all the probability mass. We also vary $\sigma$ from 0.8 to 1.1.

We observe the expected tradeoff. With smaller $\sigma$, our $C_\sigma$ is closer to the ideal distribution (as judged by $L_\infty$ error), but this imposes a stricter upper bound on the sample variance of $\Delta^*$ before our test can be applied, which thus results in larger minibatch sizes. Conversely, a more liberal upper bound means we avail ourselves of smaller minibatch sizes, but at the cost of a less stable derivation for $C_\sigma$.

We chose $N = 4000$, $\sigma = 1$, and $\lambda = 10$ to use in our experiments, which empirically exhibits excellent performance. This is reflected in the description of MHMINIBATCH in Algorithm 1, which assumes that we used $\sigma = 1$ but we reiterate that the choice is arbitrary so long as $0 < \sigma < \sqrt{\pi^2/3} \approx 1.814$, the standard deviation of the standard logistic distribution, since there must be some variance left over for $X_{\text{corr}}$. The choice is fixed to $\sigma = 1$ but we reiterate that the choice is arbitrary so long as $0 < \sigma < \sqrt{\pi^2/3} \approx 1.814$, the standard deviation of the standard logistic distribution, since there must be some variance left over for $X_{\text{corr}}$.

---

**Table 3: Errors ($L_\infty$) in $X_{\text{norm}} + X_{\text{corr}}$ versus $X_{\text{log}}$, with $N = 4000$ (top row) and $N = 2000$ (bottom row).**

<table>
<thead>
<tr>
<th>$N = 4000$</th>
<th>$\sigma = 0.8$</th>
<th>$N = 4000$</th>
<th>$\sigma = 0.9$</th>
<th>$N = 4000$</th>
<th>$\sigma = 1.0$</th>
<th>$N = 4000$</th>
<th>$\sigma = 1.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>$L_\infty$ error</td>
<td>$\lambda$</td>
<td>$L_\infty$ error</td>
<td>$\lambda$</td>
<td>$L_\infty$ error</td>
<td>$\lambda$</td>
<td>$L_\infty$ error</td>
</tr>
<tr>
<td>100</td>
<td>2.6e-3</td>
<td>100</td>
<td>3.3e-3</td>
<td>100</td>
<td>4.3e-3</td>
<td>100</td>
<td>4.6e-3</td>
</tr>
<tr>
<td>10</td>
<td>4.0e-4</td>
<td>10</td>
<td>6.4e-4</td>
<td>10</td>
<td>1.3e-3</td>
<td>10</td>
<td>7.5e-3</td>
</tr>
<tr>
<td>1</td>
<td>6.7e-5</td>
<td>1</td>
<td>1.6e-4</td>
<td>1</td>
<td>1.1e-3</td>
<td>1</td>
<td>1.3e-2</td>
</tr>
<tr>
<td>0.1</td>
<td>1.4e-5</td>
<td>0.1</td>
<td>1.3e-4</td>
<td>0.1</td>
<td>2.0e-3</td>
<td>0.1</td>
<td>2.4e-2</td>
</tr>
<tr>
<td>0.01</td>
<td>5.0e-6</td>
<td>0.01</td>
<td>2.7e-4</td>
<td>0.01</td>
<td>3.6e-3</td>
<td>0.01</td>
<td>2.4e-2</td>
</tr>
</tbody>
</table>

**Table 4: Gaussian Mixture Model statistics ($\pm$ one standard deviation over 10 trials).**

<table>
<thead>
<tr>
<th>Metric/Method</th>
<th>MHMINIBATCH</th>
<th>AUSTEREMH(C)</th>
<th>MHSUBLEHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equation 39</td>
<td>-1307.0 ± 229.5</td>
<td>-1386.9 ± 322.4</td>
<td>-1295.1 ± 278.0</td>
</tr>
<tr>
<td>Chi-Squared</td>
<td>4502.3 ± 1821.8</td>
<td>5216.9 ± 3315.8</td>
<td>3462.3 ± 1519.5</td>
</tr>
</tbody>
</table>
C.2 GAUSSIAN MIXTURE MODEL EXPERIMENT (SECTION 6.1)

C.2.1 Grid Search
For the Gaussian mixture experiment, we use the conservative method from [Korattikara et al., 2014], which avoids the need for recomputing log likelihoods of each data point each iteration by choosing baseline minibatch sizes $m$ and per-test thresholds $\epsilon$ beforehand, and then using those values for the entirety of the trials. We experimented with the following values, which are similar to the values reported in [Korattikara et al., 2014]:

- $\epsilon \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2\}$
- $m \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$

and chose the $(m, \epsilon)$ pairing which resulted in the lowest expected data usage given a selected upper bound on the error. Through personal communication with Korattikara et al. [2014], we were able to use their same code to compute expected data usage and errors.

The main difference between AUSTERE MH($C$) and AUSTERE MH($NC$) is that the latter needs to run a grid search each iteration (i.e. after each time it makes an accept/reject decision for one sample $\theta_t$). We use the same $\epsilon$ and $m$ candidates above for AUSTERE MH($NC$).

C.2.2 Gaussian Mixture Model Metrics
We discretize the posterior coordinates into bins with respect to the two components of $\theta$. The probability $P_i$ of a sample falling into bin $i$ is the integral of the true posterior over the bin’s area. A single sample should therefore be multinomial with distribution $P$, and a set of $n$ (ideally independent) samples is $\text{Multinomial}(P, n)$. This distribution is simple and we can use it to measure the quality of the samples rather than use general purpose tests like KL-divergence or likelihood-ratio, which are problematic with zero counts.

For large $n$, the per-bin distributions are approximated by Poissons with parameter $\lambda_j = P_j n$. Given samples $\{\theta_1, \ldots, \theta_T\}$, let $c_j$ denote the number of individual samples $\theta_i$ that fall in bin $j$ out of $N_{\text{bins}}$ total. We have

$$\log p(c_1, \ldots, c_{N_{\text{bins}}} | P_1, \ldots, P_{N_{\text{bins}}}) = \sum_{j=1}^{N_{\text{bins}}} c_j \log(nP_j) - nP_j - \log(\Gamma(c_j + 1)).$$

(39)

Table 4 shows the likelihoods. To facilitate interpretation we perform significance tests using Chi-Squared distribution (also in Table 4). The table provides the mean likelihood value and mean Chi-Squared test statistics value as well as their standard deviations. Our likelihood values lies between [Korattikara et al., 2014] and [Bardenet et al., 2014], but we note that we are not aiming to optimize the likelihood values or the Chi-Squared statistics. We use these values to show the extent of correctness.

C.3 LOGISTIC REGRESSION EXPERIMENT (SECTION 6.2)

Figure 5 shows the histograms for the four methods on one representative trial of MNIST-13k, indicating similar relative performance of the four methods as in Figure 4 (which uses MNIST-100k). In particular, MHMINIBATCH exhibits a shorter-tailed distribution and consumes nearly an order of magnitude fewer data points compared to AUSTERE MH($NC$), the next-best method; see Table 2 for details.

Next, we investigate the impact of the step size $\sigma$ for the random walk proposers with covariance matrix $\sigma I$. Note that $I$ is $784 \times 784$ as we did not perform any downsampling or data preprocessing other than rescaling the pixel values to lie in $[0, 1]$.

For this, we use the larger dataset MNIST-100k, and test with $\sigma \in \{0.005, 0.01, 0.05\}$. We keep other parameters consistent with the experiments in Section 6.2, in particular, keeping the initial minibatch size $m = 100$, which is also the amount the minibatch increments by if we need more data. Figure 6 indicates minibatch histograms (again, using the log-log scale) for one trial of MHMINIBATCH using each of the step sizes. We observe that by tuning

\footnote{AUSTERE MH($NC$) is used in Section 6.2.}
Figure 5: Minibatch sizes for a representative trial of logistic regression on MNIST-13k (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.

Figure 6: Effect of changing the proposal step size $\sigma$ for MHMINIBATCH.

MHMINIBATCH, we are able to adjust the average number of data points in a minibatch across a wide range of values. Here, the smallest step size results in an average of just 116.1 data points per minibatch, while increasing to $\sigma = 0.05$ (the step size used for MNIST-13k) results in an average of 2215.6. This relative trend is also present for both AUSTERE MH variants and MHSUBLHD.

Table 5 indicates the relevant parameter settings for the logistic regression experiments. Unless otherwise stated, values apply to all methods tested. For values from [Korattikara et al., 2014] or [Bardenet et al., 2014], we use their notation ($\Delta^*, m, \epsilon, \gamma, p, \text{and} \delta$) to be consistent.
Table 5: Parameters for the logistic regression experiments.

<table>
<thead>
<tr>
<th>Value</th>
<th>MNIST-13k</th>
<th>MNIST-100k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature $K$</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Number of samples $T$</td>
<td>5000</td>
<td>3000</td>
</tr>
<tr>
<td>Number of trials</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Step size $\sigma$ for random walk proposer with covariance $\sigma I$</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>MHMINIBATCH and MHSUBLHD minibatch size $m$</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>AUSTERE MH(C) chosen $\Delta^*$ bound</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>AUSTERE MH(C) minibatch size $m$ from grid search</td>
<td>450</td>
<td>300</td>
</tr>
<tr>
<td>AUSTERE MH(C) per-test threshold $\epsilon$ from grid search</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>AUSTERE MH(NC) chosen $\Delta^*$ bound</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>MHSUBLHD $\gamma$</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>MHSUBLHD $p$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MHSUBLHD $\delta$</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>