Supplementary Material

A 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 $q(\theta' \mid \theta)$ and $p(\theta \mid x_1, \ldots, x_N)$ have variance $1/N$. If we denote Φ 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 (25) 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

$$\log u_0 = N(\theta' - \theta)\frac{1}{b} \cdot E\left[\sum_{i=1}^{b} x_i - \theta - \frac{\theta' - \theta}{2}\right] \quad (24)$$

$$\log u_0 = -N(\theta' - \theta)\left(\theta - 0.5 + \frac{\theta' - \theta}{2}\right). \quad (25)$$

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 of x_i. This turns out to be

$$\begin{aligned} &\Lambda^*(\theta, \theta') \not\approx \psi(u_0, \theta, \theta') \\ \iff &N(\theta' - \theta) \cdot \frac{1}{b}\sum_{i=1}^{b} x_i - \theta - \frac{\theta' - \theta}{2} \not\approx \log u_0 \\ \iff &\frac{1}{b}\sum_{i=1}^{b} x_i - \left(\theta + \frac{\theta' - \theta}{2} + \frac{\log u_0}{N(\theta' - \theta)}\right) \not\approx 0 \\ \iff &\frac{1}{b}\sum_{i=1}^{b} x_i - 0.5 \not\approx 0. \end{aligned} \quad (26)$$

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 θ and θ' guarantee that

$$\log u_0 \in -[0.5, 1][0.75, 1.5] = [-1.5, -0.375]. \quad (27)$$

Next, consider the range of u values near u_0:

$$\log u \in \log u_0 + [-0.5, 0.375]. \quad (28)$$

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:

$$\frac{1}{b}\sum_{i=1}^{b} x_i - 0.5 \quad \not\approx \quad \frac{\log u/u_0}{N(\theta' - \theta)} \quad (29)$$

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[6] that $\theta' - \theta > 0$:

$$\begin{aligned} \min\{\text{RHS}\} &= \frac{-0.5}{\sqrt{N} \cdot 0.5} = -\frac{1}{\sqrt{N}} \\ \text{and} \quad \max\{\text{RHS}\} &= \frac{0.375}{\sqrt{N} \cdot 0.5} = \frac{0.75}{\sqrt{N}} \end{aligned} \quad (30)$$

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 θ, θ' and u are drawn independently and so the probability of the conjunction of these events is $c = p_0 p_1 p_2 = 0.0085$.

B Proof of Lemma 3

The following bound is given immediately after Corollary 2 from Novak [2005]:

$$\begin{aligned} -6.4E|X|^3 - 2E|X| &\le \sup_x |\Pr(t < x) - \Phi(x)|\sqrt{n} \\ &\le 1.36E|X|^3. \end{aligned}$$

This bound applies to $x \ge 0$. Applying the bound to $-x$ when $x < 0$ and combining with $x > 0$, we obtain the weaker but unqualified bound in Equation (18).

C Proof of Lemma 4

We first observe that

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

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

$$\begin{aligned} -\epsilon = &\int_{-\infty}^{+\infty} -\epsilon R(x)dx \\ \le &\int_{-\infty}^{+\infty} (P(z - x) - Q(z - x))R(x)dx \\ \le &\int_{-\infty}^{+\infty} \epsilon R(x)dx = \epsilon, \end{aligned}$$

[6] 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.

as desired.

D Proof of Corollary 2

We apply Lemma 4 twice. First take:

$$\begin{aligned} P(y) &= \Pr(\Delta^* < y) \\ \text{and} \quad Q(y) &= \Phi\left(\frac{y-\Delta}{s_{\Delta^*}}\right) \end{aligned} \tag{31}$$

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:

$$\begin{aligned} P'(y) &= \Pr(\Delta^* + X_{\text{nc}} < y) \\ \text{and} \quad Q'(y) &= \Phi\left(y-\Delta\right) \end{aligned} \tag{32}$$

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

$$\begin{aligned} P''(y) &= \Pr(\Delta^* + X_{\text{nc}} + X_{\text{corr}} < y) \\ \text{and} \quad Q''(y) &= S\left(y-\Delta\right) \end{aligned} \tag{33}$$

Both steps preserve the error bound $\epsilon(\theta, \theta', b)$. Finally $S(y-\Delta)$ is a logistic CDF centered at Δ, 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 $\Pr(\Delta^* + X_{\text{nc}} + X_{\text{corr}} > 0)$ differs from the exact test $\Pr(\Delta + X_{\log} > 0)$ by at most ϵ.

E 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 μ_i denote the i^{th} moment, and b_i denote the i^{th} absolute moment of X. If Cramer's condition holds:

$$\lim_{t\to\infty} \sup |E(\exp(itX))| < 1, \tag{34}$$

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 $\hat{X}$ and with t defined as in Lemma 3. If X satisfies Cramer's condition, then*

$$\sup_x \left| \Pr(t < x) - G\left(x, \frac{\mu_3}{b_2^{3/2}}\right) \right| \le \frac{c(\epsilon, b_2, b_3, b_4, b_{4+\epsilon})}{n}$$

where

$$G_n(x,y) = \Phi(x) + \frac{y(2x^2+1)}{6\sqrt{n}}\Phi'(x). \tag{35}$$

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:

$$G_n(x,y) = \Phi(x) + \frac{y}{6\sqrt{n}}(2\phi''(x) + 3\phi(x)) \tag{36}$$

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 σ between Δ^* and the variance needed by X_{corr} using X_{nc}. This same method works when we wish to estimate the error in Δ^* 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 σ' increases the variance of all three functions to $1+\sigma'$. Thus we add X_{nc} as per Algorithm 1 preserving the limit in Equation (36).

The deconvolution approach can be used to construct a correction variable X_{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 $\frac{y}{6\sqrt{n}}$ in Equation (36), its enough to consider a function $g(x,y')$ parametrized by $y' = \frac{y}{6\sqrt{n}}$. 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.

F NIPS 2016 Submission Statement

We previously submitted an early, incomplete draft of this paper to NIPS 2016. The current manuscript has been substantially improved since that submission. We have made the following changes:

1. The (incomplete) analysis in the NIPS paper has been substantially replaced. The asymptotic CLT arguments in the earlier paper have been replaced with explicit quantitative bounds. There were several gaps in the earlier analysis: the effects of addition of random variables on CDF errors and the construction and error bounds of the correction distribution which are key parts of the present paper have been added.

2. Several recent relevant references have been added, and we have written clearer statements about the contribution of the present work. These include [Bardenet et al., 2015] and the Barker test [Barker, 1965].

3. Some reviewers mentioned that our deconvolution approach to determine the correction distribution was not unique and thus ill-defined. We added detail on how the correction distribution is computed (see Section 4). While the distribution is not unique, we regularize the inversion to obtain a concrete distribution with very low error (close to single-point machine precision).

4. The earlier draft had an inconsistency in its statement about constant minibatch size. We clarify that this is true only in expected value.

5. Our experiments now include new comparisons with a baseline from [Bardenet et al., 2014] (in addition to [Korattikara et al., 2014], which we had earlier). Furthermore, each algorithm now runs MCMC sampling on the *same* distribution. Previously, we ran our distribution at a higher temperature but kept the algorithm from [Korattikara et al., 2014] running on the distribution at temperature $T = 1$. In this set of experiments, we tuned hyperparameters of the other algorithms to yield best results. In the Gaussian mixture model scenario (Section 6.1), we provide likelihoods and test statistics to measure the accuracy of the samples.

6. We improved Algorithm 1 so that we show explicitly when we compute the moments, and why we now only need one distribution C_σ for $\sigma = 1$ due to the extra X_{nc} variable.

7. Finally, we made minor revisions addressing: differences between the Barker function vs. the original MH test, and sampling with vs. without replacement.

We estimate the degree of overlap between the papers in mathematics (equations) at less than 10%, in overall text at less than 30%. The experiments are similar but were also redone at larger scale in this draft.