1 Introduction

Here we document the technical details behind GrowthBook power calculations and minimum detectable effect (MDE) calculations for both frequentist (Section 2) and Bayesian (Section 3) engines.

2 Frequentist power

2.1 Power definition

Below we describe technical details of our implementation. First we start with the definition of power.

Definition 1. Power is the probability of a statistically significant result.

We use the terms below throughout. Define:

- the false positive rate as α (GrowthBook default is $\alpha = 0.05$).
- the critical values $Z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $Z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ where Φ^{-1} is the inverse CDF of the standard normal distribution.
- the true relative treatment effect as Δ, its estimate as $\hat{\Delta}$ and its estimated standard error as $\hat{\sigma}_\Delta$. Note that as the sample size n increases, $\hat{\sigma}_\Delta$ decreases by a factor of $1/\sqrt{n}$.

We make the following assumptions:

- equal sample sizes across control and treatment variations;
- equal variance across control and treatment variations;
- observations across users are independent and identically distributed;
- all metrics have finite variance; and
- you are running a two-sample t-test. If in practice you use CUPED, your power will be higher.

2.2 Frequentist power

For a 1-sided test, the power is

$$\pi = P\left(\frac{\hat{\Delta}}{\hat{\sigma}_\Delta} > Z_{1-\alpha}\right) = P\left(\frac{\hat{\Delta}-\Delta}{\hat{\sigma}_\Delta} > Z_{1-\alpha} - \frac{\Delta}{\hat{\sigma}_\Delta}\right) = 1 - \Phi\left(Z_{1-\alpha} - \frac{\Delta}{\hat{\sigma}_\Delta}\right). \tag{1}$$

For a 2-sided test (all GrowthBook tests are 2-sided), power is composed of the probability of a statistically significant positive result and a statistically significant negative result. Using the same algebra as in Equation 1 (except using $Z_{1-0.5\alpha}$ for the critical value), the probability of a statistically significant positive result is

$$\pi_{pos} = 1 - \Phi\left(Z_{1-\alpha/2} - \frac{\Delta}{\hat{\sigma}_\Delta}\right). \tag{2}$$

The probability of a statistically significant negative result is

$$\pi_{neg} = P\left(\frac{\hat{\Delta}}{\hat{\sigma}_\Delta} < Z_{\alpha/2}\right) = P\left(\frac{\hat{\Delta}-\Delta}{\hat{\sigma}_\Delta} < Z_{\alpha/2} - \frac{\Delta}{\hat{\sigma}_\Delta}\right) = \Phi\left(Z_{\alpha/2} - \frac{\Delta}{\hat{\sigma}_\Delta}\right). \tag{3}$$

For a 2-sided test, the power equals

$$\pi = 1 - \Phi\left(Z_{1-\alpha/2} - \frac{\Delta}{\hat{\sigma}_\Delta}\right) + \Phi(Z_{\alpha/2} - \frac{\Delta}{\hat{\sigma}_\Delta}). \tag{4}$$

2.3 Frequentist minimum detectable effect

Some customers want to know what effect size is required to produce at least π power.

Definition 2. The **Minimum detectable effect** is the smallest Δ for which nominal power (e.g., 80%) is achieved.

Below we describe commonly used MDE calculations. For a 1-sided test there is a closed form solution for the MDE. Solving Equation 1 for Δ produces

$$\text{MDE} = \hat{\sigma}_\Delta\left(\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\pi)\right). \tag{5}$$

In the 2-sided case there is no closed form solution. Often in practice the MDE is defined as the solution to inverting Equation 2. This ignores the negligible term in Equation 3, and produces power estimates very close to π:

$$\text{MDE}_{\text{two-sided}} = \hat{\sigma}_\Delta\left(\Phi^{-1}(1-\alpha/2) - \Phi^{-1}(1-\pi)\right). \tag{6}$$

This approach works when effects are defined on the absolute scale, where the uncertainty of effect estimate does not depend upon the true absolute effect. For relative inference, this does not hold. Define Δ_{Abs} as the absolute effect. Define μ_A as the population mean of variation A and σ^2 as the population variance. For variation B analogously define

μ_B; recall that we assume equal variance across treatment arms. Define N as the per-variation sample size. Define the sample counterparts as ($\hat{\mu}_A$, $\hat{\sigma}^2_A$, $\hat{\mu}_B$, and $\hat{\sigma}^2_B$). Then the variance of the sample lift is

$$\hat{\sigma}^2_{\Delta} = \frac{\sigma^2}{N}\frac{1}{\mu_A^2} + \frac{\sigma^2}{N} * \frac{\mu_B^2}{\mu_A^4} \tag{7}$$

$$= \frac{\sigma^2}{N}\frac{1}{\mu_A^2} + \frac{\sigma^2}{N} * \frac{(\mu_A + \Delta_{Abs})^2}{\mu_A^4}. \tag{8}$$

Therefore, when inverting the power formula above to find the minimum Δ that produces at least 80% power, the uncertainty term $\hat{\sigma}_{\Delta}$ changes as Δ changes. To find the MDE we solve for the equation below, where we make explicit the dependence of $\hat{\sigma}_{\Delta}$ on Δ:

$$\frac{\Delta}{\hat{\sigma}_{\Delta}(\Delta)} = \Phi^{-1}\left(1 - \alpha/2\right) - \Phi^{-1}(1 - \pi).$$

Define the constant $k = \Phi^{-1}\left(1 - \alpha/2\right) - \Phi^{-1}(1 - \pi)$. We solve for μ_B in:

$$\frac{(\mu_B - \mu_A)/\mu_A}{\sqrt{\text{Var}(\hat{\Delta})}} = k \iff (\mu_B - \mu_A)^2 = k^2\mu_A^2\text{Var}(\hat{\Delta}) = k^2\mu_A^2\left(\frac{\sigma^2}{N}\frac{1}{\mu_A^2} + \frac{\sigma^2}{N} * \frac{\mu_B^2}{\mu_A^4}\right).$$

Rearranging terms shows that

$$\mu_B^2\left(1 - \frac{\sigma^2}{N}\frac{k^2}{\mu_A^2}\right) + \mu_B\left(-2\mu_A\right) + \left(\mu_A^2 - k^2\frac{\sigma^2}{N}\right) = 0.$$

This is quadratic in μ_B and has solution

$$\mu_B = \frac{2\mu_A \pm \sqrt{4\mu_A^2 - 4\left(1 - \frac{\sigma^2}{N}\frac{k^2}{\mu_A^2}\right)\left(\mu_A^2 - k^2\frac{\sigma^2}{N}\right)}}{2\left(1 - \frac{\sigma^2}{N}\frac{k^2}{\mu_A^2}\right)} = \frac{\mu_A \pm \sqrt{\mu_A^2 - \left(1 - \frac{\sigma^2}{N}\frac{k^2}{\mu_A^2}\right)\left(\mu_A^2 - k^2\frac{\sigma^2}{N}\right)}}{\left(1 - \frac{\sigma^2}{N}\frac{k^2}{\mu_A^2}\right)}.$$

The discriminant reduces to

$$k^2 * \frac{\sigma^2}{N}\left(2 - \frac{\sigma^2}{N} * \frac{k^2}{\mu_A^2}\right).$$

so a solution for μ_B exists if and only if

$$2 - \frac{\sigma^2}{N} * \frac{k^2}{\mu_A^2} > 0 \iff 2 > \frac{\sigma^2}{N} * \frac{k^2}{\mu_A^2} \iff N > \frac{\sigma^2 k^2}{2\mu_A^2}. \tag{9}$$

Similarly, the MDE returned can be negative if the denominator is negative, which is nonsensical. We return cases only where the denominator is positive, which occurs if and only if:

$$\left(1 - \frac{\sigma^2}{N}\frac{k^2}{\mu_A^2}\right) > 0 \iff \left(1 - \frac{\sigma^2}{N}\frac{k^2}{\mu_A^2}\right) > 0 \iff N > \frac{\sigma^2 k^2}{\mu_A^2}. \tag{10}$$

The condition in Equation 10 is stricter than the condition in Equation 9.

In summary, there will be some combinations of (μ_A, σ_2) where the MDE does not exist for a given N. If $\alpha = 0.05$ and $\pi = 0.8$, then $k \approx 2.8$. Therefore, a rule of thumb is that N needs to be roughly 9 times larger than the ratio of the variance to the squared mean to return an MDE. In these cases, N needs to be increased.

2.4 Sequential testing

To estimate power under sequential testing, we adjust the variance term $\hat{\sigma}_{\delta}$ to account for sequential testing, and then input this adjusted variance into our power formula. We assume that you look at the data only once, so our power estimate below is a lower bound for the actual power under sequential testing. Otherwise we would have to make assumptions about the temporal correlation of the data generating process.

In Sequential testing we construct confidence intervals as

$$\hat{\Delta} \pm \hat{\sigma} * \sqrt{N} * \sqrt{\frac{2(N\rho^2 + 1)}{N^2\rho^2}\log\left(\frac{\sqrt{N\rho^2 + 1}}{\alpha}\right)}$$

where

$$\rho = \sqrt{\frac{-2\text{log}(\alpha) + \log(-2\text{log}(\alpha) + 1)}{N^*}}$$

and $N^\star$ is a tuning parameter. This approach relies upon asymptotic normality. For power analysis we rewrite the confidence interval as

$$\begin{aligned}&\hat{\Delta} \pm \hat{\sigma} * \sqrt{N} * \sqrt{\frac{2(N\rho^2+1)}{N^2\rho^2}\log\left(\frac{\sqrt{N\rho^2+1}}{\alpha}\right)\frac{Z_{1-\alpha/2}}{Z_{1-\alpha/2}}} \\ &= \hat{\Delta_r} \pm \tilde{\sigma} Z_{1-\alpha/2}\end{aligned}$$

where

$\tilde{\sigma} = \hat{\sigma} * \sqrt{N}\sqrt{\frac{2(N\rho^2+1)}{N^2\rho^2}\log\left(\frac{\sqrt{N\rho^2+1}}{\alpha}\right)\frac{1}{Z_{1-\alpha/2}}}$.

We use power analysis described above, except we substitute $\tilde{\sigma}^2$ for $\hat{\sigma}^2_\Delta$.

3 Bayesian power and minimum detectable effect

3.1 Bayesian power

For Bayesian power analysis, we let users specify the prior distribution of the treatment effect. We then estimate Bayesian power, which is the probability that the $(1-\alpha)$ credible interval does not contain 0.

We assume a conjugate normal-normal model, as follows:

$$\begin{aligned}\Delta &\sim \mathcal{N}\left(\mu_{prior}, \sigma^2_{prior}\right) \\ \hat{\Delta}|\Delta &\sim \mathcal{N}\left(\Delta, \hat{\sigma}^2_\Delta\right).\end{aligned}$$

In words, the customer specifies a normal prior for the treatment effect, and conditional upon the treatment effect, the estimated effect is normally distributed. The normal prior has several advantages, including: 1) bell-shaped distribution around the prior mean, so that extreme estimates will be shrunk more towards the prior than moderate estimates; 2) the ability to specify two moments, which is often the right amount of information for a prior; and 3) simplicity. The conditional normality of the effect estimate is motivated by the central limit theorem.

We use the normal distribution below to approximate the posterior:

$$\begin{aligned}\Delta|\hat{\Delta} &\sim \mathcal{N}\left(\Omega^{-1}\omega, \Omega^{-1}\right) \\ \Omega &= 1/\sigma^2_{prior} + 1/\hat{\sigma}^2_\Delta \\ \omega &= \mu_{prior}/\sigma^2_{prior} + \hat{\Delta}/\hat{\sigma}^2_\Delta.\end{aligned}$$

This is an approximation to the posterior because Δ affects $\hat{\sigma}^2_\Delta$. We tested this approximation through extensive simulations, and found it had comparable coverage and mean squared error to a posterior distribution empirically sampled using Metropolis Hastings [Chib and Greenberg, 1995].

We define rejection as the $100(1-\alpha)\%$ confidence interval not containing zero. For our posterior approximation, this occurs if the posterior mean for $\Delta|\hat{\Delta}$ (i.e., $\Omega^{-1}\omega$) divided by its posterior standard deviation $\left(\text{i.e., } \sqrt{\Omega^{-1}}\right)$ is beyond the the appropriate critical threshold $Z^\star$ (e.g., $\Phi^{-1}(0.975)$ for $\alpha = 0.05$).

Inside of a Bayesian framework, it can help to permit the case where the prior model is misspecified. That is, the prior specified by the customer differs from the true prior that generates the treatment effect. We permit misspecification of the prior for Δ, as we assume that the true data generating process (DGP) is $\Delta \sim \mathcal{N}\left(\mu_\star, \sigma^2_\star\right)$, while the specified DGP has $\Delta \sim \mathcal{N}\left(\mu_{prior}, \sigma^2_{prior}\right)$. We assume the prior is specified on the relative scale.

In derivations below we use the marginal distribution of $\hat{\Delta}$, which we find using its moment generating function:

$$\begin{aligned}E\left[\exp^{t\hat{\Delta}}\right] &= E_\Delta\left[E\left[\exp^{t\hat{\Delta}}|\Delta\right]\right] \\ &= E_\Delta\left[\exp^{t\Delta + t^2\hat{\sigma}^2_\Delta}\right] \\ &= \exp^{t^2\hat{\sigma}^2_\Delta} E_\Delta\left[\exp^{t\Delta}\right] \\ &= \exp^{t^2\hat{\sigma}^2_\Delta}\exp^{t\mu^\star + t^2\sigma^2_\star} \\ &\sim \mathcal{N}\left(\mu^\star, \sigma^2_\star + \hat{\sigma}^2_\Delta\right).\end{aligned}$$

For a 2-sided test the probability of rejection is

$$
\begin{aligned}
&P\left(\left|\frac{\Omega^{-1}\omega}{\Omega^{-1/2}}\right| > Z_{1-\alpha/2}\right) \\
&= P\left(\left|\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{-1/2}\left(\frac{\mu_{prior}}{\sigma^2_{prior}} + \frac{\hat{\Delta}}{\hat{\sigma}^2_{\Delta}}\right)\right| > Z_{1-\alpha/2}\right) \\
&= P\left(\left|\left(\frac{\mu_{prior}}{\sigma^2_{prior}} + \frac{\hat{\Delta}}{\hat{\sigma}^2_{\Delta}}\right)\right| > \left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2}\right) \\
&= P\left(\left(\frac{\mu_{prior}}{\sigma^2_{prior}} + \frac{\hat{\Delta}}{\hat{\sigma}^2_{\Delta}}\right) > \left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2}\right) + P\left(\left(\frac{\mu_{prior}}{\sigma^2_{prior}} + \frac{\hat{\Delta}}{\hat{\sigma}^2_{\Delta}}\right) < -\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2}\right) \\
&= P\left(\hat{\Delta} > \hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} - \frac{\mu_{prior}}{\sigma^2_{prior}}\right]\right) + P\left(\hat{\Delta} < -\hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} + \frac{\mu_{prior}}{\sigma^2_{prior}}\right]\right) \\
&= P\left(\frac{\hat{\Delta} - \mu_{\star}}{\sqrt{\hat{\sigma}^2_{\Delta} + \sigma^2_{\star}}} > \frac{\hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} - \frac{\mu_{prior}}{\sigma^2_{prior}}\right] - \mu_{\star}}{\sqrt{\hat{\sigma}^2_{\Delta} + \sigma^2_{\star}}}\right) \\
&+ P\left(\frac{\hat{\Delta} - \mu_{\star}}{\sqrt{\hat{\sigma}^2_{\Delta} + \sigma^2_{\star}}} < \frac{-\hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} + \frac{\mu_{prior}}{\sigma^2_{prior}}\right] - \mu_{\star}}{\sqrt{\hat{\sigma}^2_{\Delta} + \sigma^2_{\star}}}\right) \\
&= 1 - \Phi\left(\frac{\hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} - \frac{\mu_{prior}}{\sigma^2_{prior}}\right] - \mu_{\star}}{\sqrt{\hat{\sigma}^2_{\Delta} + \sigma^2_{\star}}}\right) + \Phi\left(\frac{-\hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} + \frac{\mu_{prior}}{\sigma^2_{prior}}\right] - \mu_{\star}}{\sqrt{\hat{\sigma}^2_{\Delta} + \sigma^2_{\star}}}\right). \quad (11)
\end{aligned}
$$

In practice GrowthBook assumes there is a true fixed effect size, i.e., the variance of the data generating process $\sigma^2_{\star}$ equals 0, and $\mu_{\star} = \Delta$, so two-sided power is

$$
\pi = 1 - \Phi\left(\frac{\hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} - \frac{\mu_{prior}}{\sigma^2_{prior}}\right] - \Delta}{\sqrt{\hat{\sigma}^2_{\Delta}}}\right) + \Phi\left(\frac{-\hat{\sigma}^2_{\Delta}\left[\left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\hat{\sigma}^2_{\Delta}}\right)^{1/2} Z_{1-\alpha/2} - \frac{\mu_{prior}}{\sigma^2_{prior}}\right] - \Delta}{\sqrt{\hat{\sigma}^2_{\Delta}}}\right). \quad (12)
$$

We assume that $\hat{\sigma}^2_{\Delta} = 0$ for simplicity and because large values of $\hat{\sigma}^2_{\Delta}$ can result in negative MDEs (see Section 3.2). If the prior variance σ^2_{prior} equals infinity then Equation 12 reduces to Equation 4.

3.2 Bayesian MDE

MDEs are not well defined in the Bayesian literature. We provide MDEs in Bayesian power analysis for customers that are used to conceptualizing MDEs and want to be able to leverage prior information in their analysis.

We could define the MDE as the minimum value of $\mu_{\star}$ such that Equation 11 achieves at least π power. This definition is Bayesian in that it permits uncertainty in the parameters in the data generating process. However, if $\sigma^2_{\star}$ is large, then there are some combinations of parameters where the MDE can be negative. That is, negative values of μ_{star} result in power being at least π. Usually the inferential focus is the true treatment effect for the experiment (Δ), not the population mean from which Δ is just one realization ($\mu_{\star}$), so we set $\sigma^2_{\star} = 0$ and consequently, $\Delta = \mu_{\star}$. This is why in practice we frame our Bayesian MDE as, "given our prior beliefs and the data generating process, what is the probability we can detect an effect of size Δ?", where Δ is a fixed number.

Another subtlety is that for a fixed sample size, Equation 12 can be *decreasing* in effect size, illustrated by Figure 1.

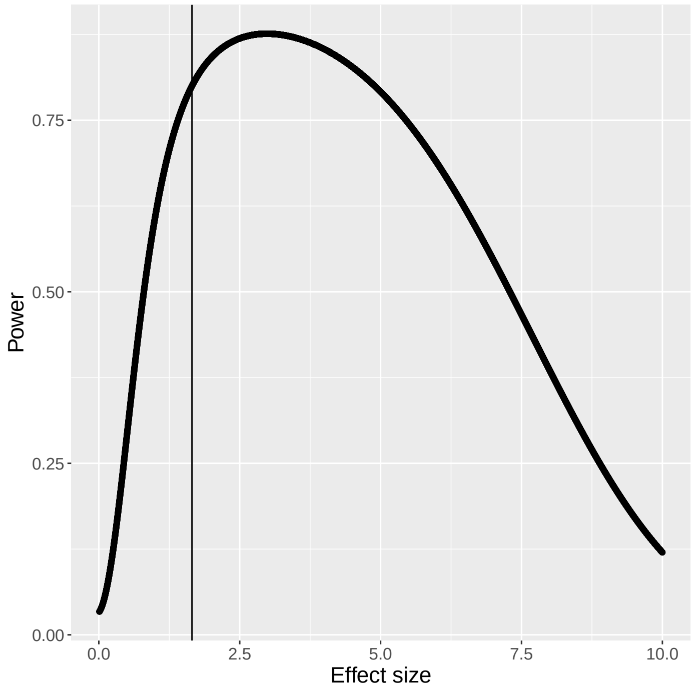

Figure 1: Bayesian power vs effect size for the case where the group sample sizes are 1500, the data mean is 0.1, the data variance is 0.5, the specified prior mean μ_{prior} is 0.1, the specified prior variance σ^2_{prior} is 0.3, and the variance of the data generating process $\sigma^2_\star$ is 0. The vertical line indicates where power equals 0.8.

Figure 1 shows the case where the group sample sizes are 1500, the data mean is 0.1, the data variance is 0.5, the specified prior mean $\mu_{prior} = 0.1$, the specified prior variance σ^2_{prior} is 0.3, and the variance of the data generating process $\sigma^2_\star$ is 0. Nominal 80% power occurs at 1.653, continues to increase in effect size until the effect size is about 3, and then begins decreasing. Power decreases in effect size because the variance in Equation 8 is quadratic in effect size, and in Equation 12 the term in front of $Z_{1-\alpha/2}$ goes to infinity as Δ gets large. The coefficient in front of $Z_{1-\alpha/2}$ in Equation 4 is 1, so frequentist power is increasing in effect size in all cases. Monotonicity does hold for Bayesian power for absolute effects, where the variance is not affected by the effect size.

Because power is not monotonic in effect size, we perform a grid search across effect sizes ranging from 0 to 500%. The derivative of Equation 12 is bounded in absolute value by $2\phi(0) < 0.8$, where $\phi(.)$ is the density of the standard normal distribution.

Let the length of one grid cell equal l. We evaluate power at the points

$$\{0, l, 2l, \ldots, 5 - l, 5\}.$$

Suppose the power at the k^{th} gridpoint is π_k, $k > 0$. Because 1) the maximum slope from the midpoint to the endpoint of the cell is no greater than $2\phi(0)$; and 2) the maximum distance from where power is evaluated is $l/2$, the maximum power in $[(l-1)k, lk]$ is no greater than $\max(\pi_{k-1}, \pi_k) + \phi(0)l$. Motivated by this fact, we describe our approach in Algorithm 1.

In words, Algorithm 1 evaluates power at $\Delta = \{0, 0.001, 0.002, ...5\}$, until it finds the first element k such that $\pi(k) >= \pi - \phi(0)l$. If power exceeds this threshold, then we evaluate a finer grid across the range from $[k - l, k]$, where the grid cell length is $l‘ << l$. We find the first element (if it exists) of this finer grid where power is at least π. We return this first element as the solution if it exists; otherwise we keep searching the coarse grid.

Algorithm 1 Power grid search

1. Define l as the length between points at which power is evaluated (l=0.001 in production).
2. Define the grid of points between 0 and 5 as

$$\mathcal{G} = \{0, l, 2l, \ldots, 5 - l, 5\}.$$

3. Begin evaluating $\pi(k)$ for $k \in \mathcal{G}$.
4. If $\pi(k) < \pi - \phi(0)l$ for all k, then the MDE does not exist. Otherwise:
5. Find the first $k \in \mathcal{G}$ such that $\pi(k) >= \pi - \phi(0)l$. Define $l^{'}$ as a finer grid resolution (in production, $l^{'} = l/100$). Find the first element of the set $\{k-1, k-1+l', k-1+2l', \ldots, k-l', k\}$ such that power evaluated at that point is at least π. If no such point exists, return to the coarser grid search in Step 3.

If power exceeds $\pi + \phi(0)l'$ at any point in $[0, 5]$, then Algorithm 1 is guaranteed to detect it. In practice we use $l = 10^{-3}$ and $l' = 10^{-5}$.

References

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.