MNRAS **000**, 1–5 (0000) Preprint 20 July 2023 Compiled using MNRAS LATEX style file v3.0

# Approximating the end of nested sampling

Zixiao Hu, Will Handley

20 July 2023

**ABSTRACT**

This project develops a technique to estimate the runtime of a nested sampling run at an intermediate iteration. The known likelihood at an iteration is fitted, then extrapolated using a Gaussian model likelihood, which allows the remaining evidence and final prior volume to be calculated. We find that the method succesfully recovers the true endpoint within standard error from the beginning for near-Gaussian likelihoods, and by the halfway point for non-Gaussian likelihoods while identifying the correct order of magnitude throughout.

**Key words:** methods: data analysis – methods: statistical

## 1 INTRODUCTION

Nested sampling is a multi-purpose algorithm invented by John Skilling which simultaneously accomplishes the two tasks of Bayesian inference, model comparison and parameter estimation (Skilling 2006). It was immediately adopted for cosmology, and is now used in a wide range of physical sciences including particle physics, materials science (Ashton et al. 2022) and machine learning (Yallup et al. 2022). The core algorithm is unique for being the only algorithm which estimates high-dimensional volumes using *order statistics*, which makes high-dimensional integration feasible. It also avoids problems faced by traditional Bayesian algorithms, such as multi-modality.

Many instances of software for nested sampling exist, including popular implementations such as MultiNest (Feroz et al. 2009) and PolyChord (Handley et al. 2015) which provide flexible options for nested sampling runs, as well as post-processing packages such as anesthetic (Handley 2019) to obtain and plot sample statistics. However, currently no such software is able to estimate how long a given run should last. This is of high importance to the end user, who has no indication in the middle of a run whether it will take minutes, or weeks and months to finish; even an order of magnitude estimate is useful compared to the status quo of complete ignorance.

This project sets out a novel approach for estimating the endpoint of a nested sampling run at an intermediate stage. The idea is to predict the form of the likelihood function in the region we have yet to sample from, using the information we have already gained from our previous samples. We begin in section 2 with an overview of Bayesian inference and nested sampling. Section 3 presents the theory and methodology of making endpoint predictions, before results for both toy and real examples are shown in section 4. Finally, conclusions are made in section 5.

## 2 BACKGROUND

7.03058in We present here an overview of Bayesian inference and nested sampling; for a more detailed treatment of both subjects, we refer to Sivia & Skilling (2006) as well as Skilling's original paper.

### 2.1 Bayesian inference

Given some data $D$, Bayes' theorem tells us that the probability of some parameters $\theta$ of a model $M$ is

$$\Pr(\theta \mid D, M) = \frac{\Pr(D \mid \theta, M)\Pr(\theta \mid M)}{\Pr(D \mid M)} \tag{1}$$

Relabelling these terms,

$$P(\theta) = \frac{L(\theta)\pi(\theta)}{Z} \tag{2}$$

where $\pi(\theta)$ is the *prior* - what was known about a model's parameters before knowledge of the data, $P(\theta)$ is the *posterior* - what is known after learning the data, and $L(\theta)$ is the *likelihood* of the data given the parameters.

For parameter estimation, knowledge of $L(\theta)$ and $\pi(\theta)$ is enough, but $Z$, known as the *evidence*, is crucial for model comparison via Bayes factors. By definition, $Z$ is the normalisation factor

$$Z = \int L(\theta)\pi(\theta)\, d\theta \tag{3}$$

which is a high-dimensional integral over the parameter space. Exponential scaling with dimension means full rasterisation is impractical for most physical applications. This motivates a sampling technique which preferentially selects for regions of high posterior mass. Traditional techniques like Markov-Chain Monte Carlo (MCMC) do not sample from the prior, so it is non-trivial to estimate the evidence from posterior samples. Nested sampling sidesteps these issues by putting the evidence first, while simultaneously yielding the posterior.

### 2.2 Nested sampling

We first perform a change of variable to reduce the evidence integral to a single dimension. Define the prior mass

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\, d\theta \tag{4}$$

which allows us to rewrite the evidence as

$$Z = \int_0^1 \mathcal{L}(X)\, dX. \tag{5}$$

The integral can be calculated by ordering the prior samples by likelihood then multiplying by the associated prior mass element. Nested sampling translates this into an algorithm which incorporates sorting *de facto*, by taking a set of $n_{\rm live}$ *live points* and replacing the point with the lowest likelihood with a resampled point of higher likelihood at every iteration. A choice must be made for this sampling method e.g. MCMC. The number of live points can in principle vary throughout the run, so that it depends on the iteration $k$ i.e. $n_{\rm live} = n_k$ (Higson et al. 2019).

The set of live points will therefore occupy a prior volume which tends to zero, shrinking around the maximum of the posterior. Explicit calculation of the volume $X$ within each contour is expensive, so nested sampling makes a statistical estimate using the order statistics of the remaining points: at each iteration, the volume occupied by the live points decreases by the shrinkage ratio $t$, which has probability distribution and statistics

$$\Pr(t) = n_{\rm live} t^{n_{\rm live}-1}, \quad \mathrm{E}[\log t] = -\frac{1}{n_{\rm live}}, \quad \mathrm{Var}[\log t] = \frac{1}{n_{\rm live}^2}. \tag{6}$$

$X_i$ is the product of the shrinkage ratios

$$X_i = \prod_{k=0}^{i} t_k, \tag{7}$$

so the $X$ themselves are also random variables. It is easier to work with $\log X$, which have the means and correlations

$$\mathrm{E}[\log X_i] = -\sum_{k=0}^{i} \frac{1}{n_k}, \quad \mathrm{Cov}[\log X_i, \log X_j] = \sum_{k=0}^{\min(i,j)} \frac{1}{n_k^2}. \tag{8}$$

The evidence integral is then computed using the trapezoid rule at each step, so that

$$Z = \sum_i L_i \Delta w_i = \frac{1}{2} \sum_i L_i \left(X_{i-1} - X_{i+1}\right). \tag{9}$$

The algorithm terminates when the evidence in the live points falls below a user-specified fraction $\epsilon$ of the total accumulated evidence. Skilling's original paper estimates this remaining evidence $\Delta Z$ to have an upper bound of $\mathcal{L}_{\max} X_i$, while some use the average remaining likelihood. Once we have chosen to stop, the remaining evidence is included by killing the live points one by one without replacement, incrementing $Z$ as before.

The full nested sampling procedure is given in algorithm 1.

**Algorithm 1** Nested sampling

Start with $N$ points $\theta_1, \ldots, \theta_N$ from prior
  initialise $Z = 0$, $X_0 = 1$
**for** $i = 1, 2, \ldots, j$ **do**
  $L_i \leftarrow$ lowest of the current likelihood values
  $X_i \leftarrow \exp(-i/N)$
  $w_i \leftarrow X_{i-1} - X_i$ or $(X_{i-1} - X_{i+1})/2$
  $Z \leftarrow Z + L_i w_i$
  $L_i \leftarrow$ new point $\in \{\pi(\theta) : L(\theta) > L_i\}$.
**end for**
Increment $Z$ by $\langle L(\theta) \rangle X_j$.

### 2.3 Uncertainties and the posterior bulk

Most of the posterior is contained within a small fraction of the prior of size $X^* = e^{-D_{\rm KL}}$, where

$$D_{\rm KL} = \int P(\theta) \log \frac{P(\theta)}{\pi(\theta)} \, d\theta \tag{10}$$

**Figure 1.** Plot showing the likelihood and the posterior weights $\mathcal{L}(X)X$. The bulk of posterior mass is contained within a region of size $e^{-D_{\rm KL}}$, meaning one must iterate through at least $nD_{\rm KL}$ steps to fully explore the posterior.

**Figure 2.** Plot showing samples of $\log X$ plotted against the likelihoods of the sample points. The uncertainty increases with the number of steps taken.

is the information gain between the prior and posterior; this is visualised in figure 1. To learn anything about the posterior or calculate an accurate evidence, we must compress down to this volume.

From equation (8), it is easy to see that for a constant number of live points $n$ $\log X$ follows Poisson statistics:

$$\mathrm{E}[\log X_i] = -\frac{i}{n}, \quad \mathrm{Var}[\log X_i] = \frac{i}{n^2}. \tag{11}$$

This is illustrated in figure 2, where samples are taken from the $X$ distribution and plotted against the likelihood values. To reach the posterior bulk we must therefore take $i^* = nD_{\rm KL}$ steps, which translates to an uncertainty in $\log X^* = -D_{\rm KL} \pm \sqrt{D_{\rm KL}/n}$. The evidence is $Z = \langle \mathcal{L} \rangle X^*$, or

$$\log Z = \langle \log \mathcal{L} \rangle + \log X^*. \tag{12}$$

The first term has low error, since the only difference that errors in $X$ makes is to slightly change the region we average over. Uncertainties in $\log Z$ are therefore dominated by the $O(1/\sqrt{n})$ errors on $\log X_i$, which are much more significant than the $O(n^{-2})$ quadrature errors.

The next section will outline the methodology for making endpoint predictions.

## 3 ENDPOINT PREDICTION

We shall begin by discussing a simple "eyeball" method for making predictions that is familiar to experienced users of nested sampling, which makes a good sanity check to compare to the more principled approach.

| iteration | log Z | $d \log Z$ |
|---|---|---|
| 6000 | -34.481 | 1.317 |
| 6200 | -32.606 | 1.875 |
| 6400 | -30.700 | 1.906 |
| 6600 | -29.008 | 1.692 |
| 6800 | -27.416 | 1.591 |
| 7000 | -25.865 | 1.551 |
| 7200 | -24.449 | 1.416 |
| 7400 | -23.034 | 1.415 |
| 7600 | -21.617 | 1.417 |
| 7800 | -20.267 | 1.349 |
| 8000 | -18.950 | 1.317 |

**Table 1.** Output of accumulated log-evidence during a nested sampling run.

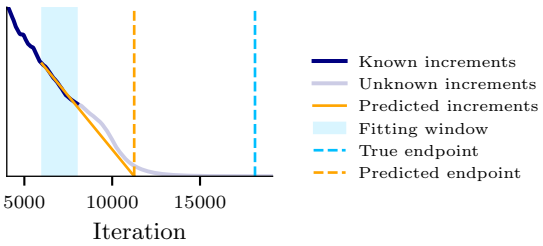



**Figure 3.** Prediction of endpoint using linear extrapolation. The estimate is much lower than the true endpoint, primarily because of the long tail in $d \log Z$.

### 3.1 Eyeball method

At an intermediate iteration of a nested sampling run, the output might look something like that shown in the first two columns of table 1. The outputs $\log Z$ appear to converge at a predictable rate, which is shown in terms the increments in column three. One could feasibly extrapolate these to predict when the increments approach zero, which gives an iteration number for convergence.

This is shown for a linear extrapolation in figure 3. The prediction is clearly an underestimate, which is mainly due to the nonlinear profile of $d \log Z$ having a long tail. However, this makes a good benchmark to beat for the more principled method using likelihood approximation.

### 3.2 Likelihood approximation method

We want to make an estimate of the number of iterations until the termination condition is met, which is when the remaining evidence falls below a fraction $\epsilon$ of the accumulated evidence. Predictions can be made ahead of time because of the live points, which are sampled from higher likelihood regions that we have not yet compressed to. It is important to note that there will always be a small region which the live points do not penetrate because they are too diffuse, which can be seen in figure 4.

If we could estimate the form of $\mathcal{L}(X)$ at the higher likelihoods in terms of some *likelihood model* $f(X, \boldsymbol{\theta})$, we would be able to express the prior volume $X_f$ which we need to compress to as

$$\Delta Z = \epsilon Z_{\text{tot}}, \tag{13}$$

$$\int_0^{X_f} f(X, \boldsymbol{\theta})\, dX = \epsilon \left( \int_0^{X_i} f(X, \boldsymbol{\theta})\, dX + Z_{\text{dead}} \right), \tag{14}$$

where $X_i$ is the volume of the iteration we have currently compressed to, and $Z_{\text{dead}}$ is the evidence we have accumulated up to this point. $X_f$ can then be identified by solving the above equation either analytically or numerically. Remembering that at each iteration $\log X$ decreases by $1/n_{\text{live}}$, the total number of iterations $N_f$ will be

$$N_f = -n_{\text{live}} \log X_f. \tag{15}$$



**Figure 4.** Plots showing the state of a nested sampling at an intermediate iteration. The live points are samples from higher likelihood regions, which allow us to see the remaining likelihood ahead of time. However, there is always a small high likelihood region where we have no samples, simply because the live points are too diffuse.

#### 3.2.1 *Model choice*

The choice of the model $f(X, \boldsymbol{\theta})$ serves two purposes; to *smooth* over the stochasticity in the live points, since $\mathcal{L}(X)$ is typically a smooth function, and more importantly to *extrapolate* into the region where we have no samples.

Making this extrapolation necessitates assumptions about the form of the likelihood in the unsampled region. It is sensible to assume it is guided by what came before it, so we fit the model on the known likelihods of the live and dead points. Apart from this, we should make the model $f(X, \boldsymbol{\theta})$ as flexible as possible to extrapolate without bias. We chose the model to be a Gaussian with freely varying parameters $\boldsymbol{\theta} = (\mathcal{L}_{\max}, d, \sigma)$:

$$f(X, \boldsymbol{\theta}) = \mathcal{L}_{\max} \exp\left( -\frac{X^{2/d}}{2\sigma^2} \right). \tag{16}$$

The choice was made because it is analytically tractable, and because many likelihoods in practice are near-Gaussian at their peaks. Allowing $d$ to vary freely increases flexibility, and reflects the fact that while the dimensionality of the posterior is fixed, the equivalent dimension for a Gaussian of the same information content can differ (Handley & Lemos 2019).

#### 3.2.2 *Endpoint formula*

Now we are equipped with the tools to make endpoint predictions. We will begin by fitting the model over the live points only. At each iteration, estimate the prior volumes of the live points to be those

which correspond to killing off each point one by one without replacement and using formula (8). Combined with the live likelihoods $\{\log \mathcal{L}_i\}$, the parameters of the model can be found by minimising the cost function

$$C^2(\boldsymbol{\theta}) = \sum_i \left|\log \mathcal{L}_i - \log f(X_i, \boldsymbol{\theta})\right|^2 \tag{17}$$

with respect to $\boldsymbol{\theta}$. A key feature is that we can analytically differentiate $C^2$ for the Gaussian model, which allows us to write the optimal values of $\log \mathcal{L}_{\max}$ and $\sigma$ in terms of the optimal values of $d$:

$$\sigma^2 = \frac{N \sum_i X_i^{4/d} - \left(\sum_i X_i^{2/d}\right)^2}{2 \sum_i \log \mathcal{L}_i \sum_i X_i^{2/d} - 2N \sum_i \log \mathcal{L}_i X_i^{2/d}} \tag{18}$$

$$\log \mathcal{L}_{\max} = \frac{1}{N} \sum_i \{\log \mathcal{L}_i\} + \frac{1}{2N\sigma^2} \sum_i X_i^{2/d} \tag{19}$$

(where $N$ here is the number of data points in we fit over), and simply minimise with respect to $d$ instead of all three variables. The result is that we are able to consistently find *global* rather than local minima in a single optimisation run. We do not need to concern ourselves with the usual difficulties of global optimisation like comparing an ensemble of initial conditions, which is orders of magnitude slower.

Once we have obtained the best-fit $\boldsymbol{\theta}$, the endpoint can be calculated in the manner described in (14);

$$\epsilon = \frac{\int_0^{X_f} \mathcal{L}_{\max} \exp\left(-X^{2/d}/2\sigma^2\right)\, dX}{\int_0^{X_i} \mathcal{L}_{\max} \exp\left(-X^{2/d}/2\sigma^2\right)\, dX + Z_{\text{dead}}}. \tag{20}$$

The integrals have the analytic solution

$$\int_0^{X_k} \mathcal{L}_{\max} \exp\left(-X^{2/d}/2\sigma^2\right)\, dX = \frac{d}{2} \cdot 2^{d/2} \cdot \sigma^d \cdot \gamma_k \tag{21}$$

where $\gamma_k = \gamma\left(d/2, X_k^{2/d}/2\sigma^2\right)$ is the lower incomplete gamma function. After taking the inverse of $\gamma$ and a few more steps of algebra, we arrive at

$$\log X_f = \frac{d}{2} \log 2 + d \log \sigma + \log \gamma^{-1}\left(d/2,\ \epsilon\left(\gamma_i + 2^{-d/2}\sigma^{-d} Z_{\text{dead}}/\mathcal{L}_{\max}\right)\right), \tag{22}$$

and $N_f$ is just $-n_{\text{live}}$ multiplied by this.

An useful feature is that the formula is robust when we are far away from the posterior bulk. In this phase, the inferred values of $\mathcal{L}_{\max}$ turn out to be very uncertain and often divergent, but the dead evidence is also negligible, so $\mathcal{L}_{\max}$ has little impact on the prediction. Once $Z_{\text{dead}}$ becomes comparable to the live evidence $\mathcal{L}_{\max}$ does begin to impact the prediction, but at this point we have reached the posterior bulk, so we expect $\mathcal{L}_{\max}$ to be well-estimated and thus contribute in a useful way.

Note that the prediction procedure should not contribute any significant time complexity to the nested sampling run, since it involves no likelihood evaluations.

#### 3.2.3 Uncertainties

Our lack of knowledge about the unsampled region means that there is always the chance that there is some huge likelihood spike at small $X$. Since there is no way to anticipate this, all uncertainty estimates are underestimates in the sense that they cannot account for this possibility; there is no estimate which will *always* include the true endpoint.

**Figure 5.** Plot showing fits of the likelihood corresponding to different draws of $\log X$. The different realisations of the likelihood lead to a set of endpoint predictions, from which a mean and variance can be obtained.

**Figure 6.** An example “wedding cake" likelihood which has plateaus on top of a Gaussian profile

We will therefore restrict ourselves to quantifying the uncertainties associated with how we have modelled the *known* data, remembering that because of our lack of knowledge, the true endpoint may fall significantly outside of this range. These uncertainties have two sources: the variance in the prior volume estimates, and the uncertainty in the least squares minimisation. Since the former is the main error in nested sampling, we expect it to dominate. As we saw in section 2.3, different sets of $\{X_i\}$ drawn from its probability distribution lead to different "maps" of what the remaining likelihood looks like, which lead to different endpoint predictions. The natural way to handle this is to sample from $\Pr(\boldsymbol{X})$ then find the best-fit $\boldsymbol{\theta}$ and endpoint for each sample, from which we can obtain the mean and variance in the estimates. An visualisation of this approach is shown in figure 5.

## 4 RESULTS

### 4.1 Toy data

Let us now apply the methodology developed above in practice. We consider the following toy examples:

(a) A spherically symmetric Gaussian, $\log \mathcal{L} = -|\boldsymbol{r}|^2/2\sigma^2$ in $d$ dimensions
(b) A discontinous "wedding cake" likelihood with Gaussian profile and parameter $\alpha_{\text{w}}$ controlling the plateau depth, as presented in Fowlie et al. (2021) and shown in figure 6.
(c) A Cauchy likelihood $\log \mathcal{L} = -\frac{1+d}{2} \log(1 + |\boldsymbol{r}|^2/\gamma^2)$.

Parameters were chosen to have large values of $D_{\mathrm{KL}}$, which are more difficult to model due to the large distance to the posterior bulk.

Nested sampling runs were made for the above likelihoods for $n_{\mathrm{live}} = 500$ (plots for different numbers of live points are not shown here, but the results are broadly the same). The state at each iteration (i.e. the likelihoods of the dead and live points) was then extracted using ANESTHETIC. At each iteration, the model is fitted over a number of points determined by the bandwidth selection procedure outlined above. For the error bars, we draw an *ensemble* of 25 $\boldsymbol{X}$ samples from the distribution (7), and find the optimal $\boldsymbol{\theta}$ and corresponding $N_f$ for each sample to get the mean and variance of the prediction.

Results are shown in figure 7. For all cases, the predictions become more accurate as the iteration increases because the samples shrink around and eventually cross the posterior bulk, leading to an increasingly representative picture of the posterior. The method works very well for the first two likelihoods, recovering the correct endpoint within standard error for the entire run. This is expected, since the model likelihood has the same form as the true likelihoods, so one expects the extrapolation made to be accurate even when we are far away from the posterior bulk.

The Cauchy likelihood shows the result when this is not the case; without samples in the bulk, a Gaussian extrapolation performs poorly. However, the correct order of magnitude is still obtained, which is a significant improvement on the status quo of complete ignorance.

### 4.2 Real data

The model was tested on real nested sampling runs for quantification of cosmological parameter tensions. Results are shown in figure 8, with the likelihood approximation method compared against the benchmark.

The likelihood model clearly performs better. For all of the chains, the true endpoint was recovered within arund standard error by the halfway point, and falls within the correct order of magnitude throughout all of the runs. In comparison, the benchmark model mostly failed to get the correct order of magnitude on the number of iterations remaining, especially near the end of each run.

Furthermore, the first four likelihoods are highly non-Gaussian, which validates the choice of a Gaussian likelihood model. For all chains, further work needs to be done to place a more accurate upper bound on the prediction uncertainty at early iterations.

## 5 CONCLUSIONS

In this project we have developed a model to estimate the endpoint of a nested sampling run, which has no precedent in the literature. This was done by predicting the form of the likelihood in the region that the samples have not reached, via fitting of a Gaussian model likelihood to the points already known to us. Assuming that the likelihood continues in a similar trajectory leads to an estimate of the remaining evidence, and hence the end iteration.

The approach taken was shown to be highly successful for near-Gaussian likelihoods, finding the true endpoint within error from the start of the run; this was expected, because the form of the model matched the likelihood. For a Cauchy likelihood as well as the real nested sampling runs that were tested, the estimates were less accurate, but still able to find the correct order of magnitude which is useful to the end user. Furthermore, the correct endpoint was able to be recovered within standard error at the halfway point of the runs.

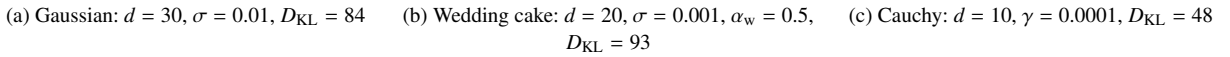
Two main refinements are recommended: the uncertainty quantification should be expanded to include the true endpoint from the beginning of the run, and the analysis should be extended to include the iteration prediction when the number of live points is allowed to vary, so that endpoints can be estimated for dynamic nested sampling.

## DATA AVAILABILITY

The data and code for the project can be found in the repository `https://github.com/zixiao-h/aeons`.

## REFERENCES

Ashton G., et al., 2022, Nature Reviews Methods Primers, 2
Feroz F., Hobson M. P., Bridges M., 2009, Monthly Notices of the Royal Astronomical Society, 398, 1601
Fowlie A., Handley W., Su L., 2021, Monthly Notices of the Royal Astronomical Society, 503, 1199
Handley W., 2019, Journal of Open Source Software, 4, 1414
Handley W., Lemos P., 2019, Physical Review D, 100
Handley W. J., Hobson M. P., Lasenby A. N., 2015, Monthly Notices of the Royal Astronomical Society, 453, 4385
Higson E., Handley W., Hobson M., Lasenby A., 2019, Statistics and Computing, 29, 891–913
Sivia D. S., Skilling J., 2006, Data Analysis: A bayesian tutorial (Oxford Science Publications). Oxford University Press
Skilling J., 2006, Bayesian Analysis, 1
Yallup D., Handley W., Hobson M., Lasenby A., Lemos P., 2022, Split personalities in Bayesian Neural Networks: the case for full marginalisation (arXiv:2205.11151)



(a) Gaussian: $d = 30$, $\sigma = 0.01$, $D_{\mathrm{KL}} = 84$ (b) Wedding cake: $d = 20$, $\sigma = 0.001$, $\alpha_{\mathrm{w}} = 0.5$, $D_{\mathrm{KL}} = 93$ (c) Cauchy: $d = 10$, $\gamma = 0.0001$, $D_{\mathrm{KL}} = 48$

**Figure 7.** Plot showing endpoint predictions for the toy likelihoods. The mean estimate is dark blue, while the shading shows the $1\sigma$ and $2\sigma$ uncertainties. For the first two likelihoods (which have Gaussian profile), the true endpoint falls within error for practically all of the run. The Cauchy likelihood is less well-estimated because it is highly non-Gaussian, with the true endpoint outside the reasonable uncertainty range but in the correct order of magnitude.

**Figure 8.** Plots showing the endpoint predictions for the benchmark model as well as the likelihood approximation model. The likelihood model visibly performs better, managing to recover the true endpoint within around standard error at the halfway point for all the chains. The first four likelihoods are highly non-Gaussian, which validates the choice of a Gaussian model.