Approximating the end of nested sampling

Zixiao Hu, Artyom Baryshnikov, Will Handley

26 September 2023

ABSTRACT

This paper develops a technique to estimate the runtime of a nested sampling run at an intermediate iteration which works for any nested sampling implementation. The known likelihood at an iteration is fitted, then extrapolated using a Gaussian model function, which allows the remaining evidence and prior volume at termination to be calculated. We find that the method succesfully recovers the true endpoint within standard error from the beginning for near-Gaussian likelihoods. For non-Gaussian likelihoods, the estimate is still the correct order of magnitude and converges to the correct answer by the end of the run.

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 (Higson et al. 2018). The core algorithm is unique for estimating volumes using *order statistics*, which makes high-dimensional integration feasible. It also avoids problems faced by traditional Bayesian algorithms, such as multi-modality.

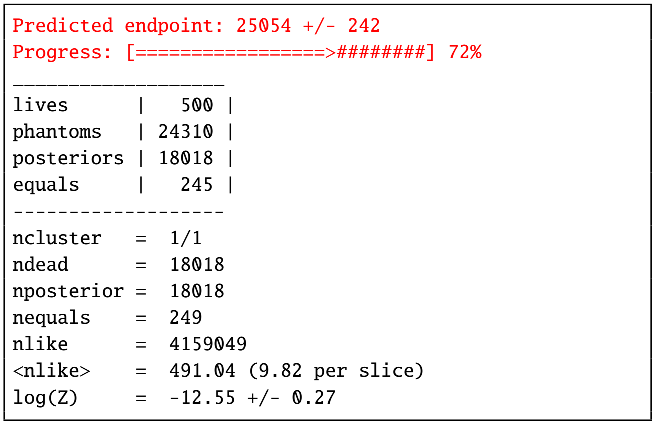
Many instances of software for nested sampling exist, including popular implementations such as MultiNest (Feroz et al. 2009) and PolyChord (Handley et al. 2015), as well as post-processing packages such as anesthetic (Handley 2019) to plot sample statistics and distributions. However, no such software is currently 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 hours, 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, an example of which is shown in Fig. 1. 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 Section 2 with an overview of nested sampling, before discussing its convergence properties in Section 3. Section 4 presents the methodology of making endpoint predictions, before results for both toy and real examples are shown in section Section 5. Finally, conclusions are made in section Section 6.

Users would like a order of magnitude runtime estimate

2 BACKGROUND

Let us begin with a brief description of the nested sampling algorithm and several adjacent concepts to establish the necessary notation.



```
Predicted endpoint: 25054 +/- 242
Progress: [=================>########] 72%
___________________
lives       |   500 |
phantoms    | 24310 |
posteriors  | 18018 |
equals      |   245 |
-------------------
ncluster   =  1/1
ndead      =  18018
nposterior =  18018
nequals    =  249
nlike      =  4159049
<nlike>    =  491.04 (9.82 per slice)
log(Z)     =  -12.55 +/- 0.27
```

Figure 1. Output from PolyChord for a typical nested sampling run. The predicted endpoint, shown in red, is calculated using the method described in this paper.

2.1 Nested sampling

For a given likelihood $\mathcal{L}(\theta)$ and prior $\pi(\theta)$, nested sampling simultaneously calculates the Bayesian evidence

$$\mathcal{Z} = \int \mathcal{L}(\theta)\, \pi(\theta)\, \mathrm{d}\theta \tag{1}$$

while producing samples of the posterior distribution

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

The algorithm operates by maintaining a set of n_{live} *live points* sampled from the prior, which can vary in number throughout the run. At each iteration, the point with the lowest likelihood is removed and added to a list of *dead points*. A new point is then drawn from the prior, subject to the constraint that it must have a higher likelihood than the latest dead point. Repeating the procedure leads to the live points shrinking around peaks in the likelihood.

The integral in Eq. (1) is then evaluated by transformation to a one-dimensional integral over the *prior volume* X, defined as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)\,\mathrm{d}X \approx \frac{1}{2}\sum_{i=1} \mathcal{L}(X_{i-1} - X_{i+1}), \tag{3}$$

where $X(\mathcal{L})$ is the fraction of the prior greater than $\mathcal{L}$. The prior volumes X_i are unknown, but can be statistically estimated as follows: one can define a *shrinkage factor* t_i at each iteration $X_i = t_i X_{i-1}$, such that

$$X_i = \prod_{k=1}^{i} t_k. \tag{4}$$

The t_i are the maximum of n_live points drawn from $[0, 1]$, so follow the distribution

$$P(t_i) = n_\mathrm{live} t_i^{n_\mathrm{live}-1}, \quad \langle \log t_i \rangle = -\frac{1}{n_\mathrm{live}}, \quad \mathrm{Var}(\log t_i) = \frac{1}{n_\mathrm{live}^2}. \tag{5}$$

The algorithm terminates when an user-specified condition is met; a popular choice is to terminate when the evidence in the live points falls below some fraction ϵ of the accumulated evidence e.g. 10^{-3}. The remaining live points are then killed off one by one without replacement and added to the evidence.

Uncertainties in the evidence are dominated by the spread in the prior volume distribution, and the simplest way to estimate them is by Monte Carlo sampling over sets of $\boldsymbol{t}$.

The time complexity of nested sampling, as analysed in Petrosyan & Handley (2022), is

$$T \propto n_\mathrm{live} \times \langle \mathcal{T}\{\mathcal{L}(\theta)\}\rangle \times \langle \mathcal{T}\{\mathrm{Impl.}\}\rangle \times \mathcal{D}_\pi\{\mathcal{P}\}. \tag{6}$$

For a given resolution of n_live, the runtime is therefore proportional to the average time for a likelihood evaluation, the implementation-dependent average time to sample a higher likelihood live point, and the information gain from prior to posterior. The first two of these are typically constant over the course of a run, so the runtime is primarily determined by the compression factor $\mathcal{D}$.

2.2 Thermodynamic connection

The thermodynamic generalisation of Bayes' theorem is given by

$$\mathcal{P}_\beta(\theta) = \frac{\mathcal{L}^\beta(\theta)\pi(\theta)}{\mathcal{Z}(\beta)}, \quad \mathcal{Z}(\beta) = \int \mathcal{L}(\theta)^\beta \pi(\theta)\,\mathrm{d}\theta, \tag{7}$$

where β is the inverse temperature $1/T$. If each θ is a microstate, then $\log\mathcal{L}$ becomes the negative energy, π becomes the density of states, and $\mathcal{Z}$ naturally the partition function. The advantage of nested sampling for computing partition functions is that it compresses at a constant thermodynamic speed by taking samples equally spaced in volume entropy, which sidesteps the problem of designing an annealing schedule (Pártay et al. 2010).

2.3 Bayesian model dimensionality

Following the thermodynamic connection, an important construct is the analogue of the heat capacity, which is the *Bayesian model dimensionality*, defined as the posterior variance of the Shannon information:

$$\frac{\tilde{d}}{2} = \int \mathcal{P}(\theta)\left(\log\frac{\mathcal{P}(\theta)}{\pi(\theta)} - \mathcal{D}_\mathrm{KL}\right)^2 \mathrm{d}\theta = \langle \mathcal{I}^2\rangle_\mathcal{P} - \langle \mathcal{I}\rangle_\mathcal{P}^2. \tag{8}$$

Just as the heat capacity is a measure of the number of degrees of freedom in a system, the Bayesian model dimensionality is a measure of the number of constrained parameters in a set of samples.

3 THE PROGRESS OF A NESTED SAMPLING RUN

We will now present several useful ways of measuring the progress of a nested sampling run, together with an assortment of results we shall use to make endpoint predictions. At a given iteration i^*, these are the current prior volume $\log X^*$ of the latest dead point, the corresponding log-likelihood $\log\mathcal{L}^*$, and the current effective temperature β^* of the samples.

3.1 Current prior volume

The conventional perspective for the progress of a run is with $\log X$ along the x-axis, where the prior volume is compressed at a steady rate until we reach the posterior mass. Defining the Kullback-Leibler divergence as

$$\mathcal{D}_\mathrm{KL} = \int \mathcal{P}(\theta)\log\frac{\mathcal{P}(\theta)}{\pi(\theta)}\,\mathrm{d}\theta, \tag{9}$$

we interpret it as the posterior averaged logarithmic *volume compression* from the prior to the posterior. Hence we can define a *posterior bulk* at below $X = e^{-\mathcal{D}_\mathrm{KL}}$ containing nearly all of the information, which gives us a measure of the volume we must compress down for the algorithm to converge.

At an intermediate iteration i_*, we only have samples of the posterior only down to the current prior volume $\log X_*$. We do have some information below this volume due to the live points, but we will see now that they do not bring us much further. If we kill off all the live points as if we were terminating, then by Eq. (5) the prior volume of the final live point $\log X_\mathrm{min}^\mathrm{live}$ has mean and variance given by

$$\mathrm{E}[\log X_\mathrm{min}^\mathrm{live}] = \mathrm{E}[\log X_*] - \sum_{k=1}^{n_\mathrm{live}} \frac{1}{n_k} \approx -\frac{i_*}{n_\mathrm{live}} - \log n_\mathrm{live} - \gamma, \tag{10}$$

$$\mathrm{Var}[\log X_\mathrm{min}^\mathrm{live}] = \mathrm{Var}[\log X_*] + \sum_{k=1}^{n_\mathrm{live}} \frac{1}{n_k^2} \approx \frac{i_*}{n_\mathrm{live}^2} + \frac{\pi^2}{6}, \tag{11}$$

where the large n_live limit is taken for the approximation to the harmonic series, γ being the Euler-Mascheroni constant.

This shows that the live points only get us a factor of $\log n_\mathrm{live}$ closer to the posterior bulk. In other words, it is not until we are around $\log n_\mathrm{live}$ away from $\log X = \mathcal{D}_\mathrm{KL}$ that the samples look anything like the posterior. One can see from Eq. (9) that the divergence increases linearly with dimension, so for large dimensionalities (such as the Planck legacy chains, which have a $\mathcal{D}_\mathrm{KL}$ of ~ 40) and typical live point numbers of ~ 1000, this does not happen until near the end of the run.

Intuitively, this is because for a sharply peaked likelihood the live points are too diffuse to land there with any significant probability for most of the run.

3.2 Current log-likelihood

We can also view the run as a function of the current log-likelihood value $\log\mathcal{L}^*$, which of course monotonically increases throughout the run. To illustrate some of the analytics, let us take the example of a Gaussian likelihood with dimension d and lengthscale σ;

$$\log\mathcal{L} = \log\mathcal{L}_\mathrm{max} - \frac{X^{2/d}}{2\sigma^2}. \tag{12}$$

$\log\mathcal{L}^*$ must increase until the posterior bulk is passed, which is located around the posterior averaged log-likelihood

$$\langle \log\mathcal{L}\rangle_\mathcal{P} = \log\mathcal{L}_\mathrm{max} - \frac{d}{2}, \quad \mathrm{Var}(\log\mathcal{L})_\mathcal{P} = \frac{d}{2}. \tag{13}$$

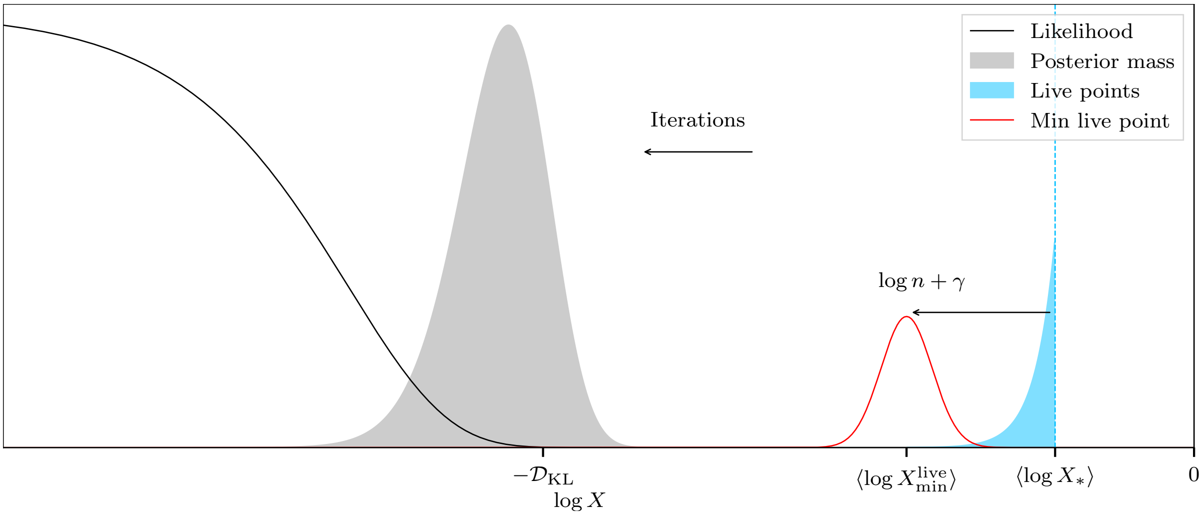


Figure 2. The distribution of the posterior mass in terms of $\log X$, the live points over the prior $\pi(\log X \mid \log X < \log X_*)$ and the smallest live point prior volume $\log X_{\mathrm{min}}^{\mathrm{live}}$ at an intermediate iteration i_*. For large values of $\mathcal{D}_{\mathrm{KL}}$ i.e. informative posteriors and/or large dimensionalities, the maximum live point is very far from the posterior bulk until the very end of the run. Note that the x-axis in this plot is $\log X$, so that the run proceeds from right to left to emphasise that the enclosed prior volume iteratively gets smaller. Plots of the sort from here onwards will be in terms of $-\log X$, where the run will more naturally proceed from rleft to right.

Given the value of the latest dead point, we can examine the distribution of the live points in likelihood. Define the normalised likelihood

$$y = \frac{\log \mathcal{L} - \log \mathcal{L}_*}{\log \mathcal{L}_{\mathrm{max}} - \log \mathcal{L}_*} \tag{14}$$

as a measure of how far a point is between the current likelihood and the maximum; $y = 0$ corresponds to $\log \mathcal{L}_*$ and $y = 1$ to $\log \mathcal{L}_{\mathrm{max}}$. It follows that the live points are distributed as

$$P(y) = \frac{d}{2}(1-y)^{\frac{d}{2}-1} \quad [0 < y < 1], \tag{15}$$

and the maximum live point has mean and variance

$$\lim_{d, n_{\mathrm{live}} \to \infty} y_{\mathrm{max}}^{\mathrm{live}} \sim \frac{2 \log n_{\mathrm{live}}}{d} \pm \sqrt{\frac{2}{3}\frac{\pi}{d}} \tag{16}$$

in the limit of large d and n_{live} (see appendix for derivation). We therefore see that just as with the prior volumes, in high dimensions the live points only bring us logarithmically further, while d pushes us linearly away.

Nested sampling as a maximiser

Previous literature has explored the potential of nested sampling algorithms as maximisers (Akrami et al. 2010; Feroz et al. 2011). The latter authors emphasised the terminatioin criterion controls the global maximum obtained.

We can use these results to show that nested sampling by itself is a poor maximiser. If we set $\log \mathcal{L}^*$ to be the termination point $\log \mathcal{L}_{\mathrm{end}}$, we can invert Eq. (14) to find that the maximum live point is located at

$$\boxed{\log \mathcal{L}_{\mathrm{max}}^{\mathrm{live}} \approx \log \mathcal{L}_{\mathrm{max}} - \frac{d}{2e} + \frac{\log n}{e} \pm \frac{\pi}{\sqrt{6}e}}. \tag{17}$$

Hence in general nested sampling will finish at a contour $d/2e$ away from the maximum log-likelihood, and the final set of n live points can get you $(\log n)/2e$ closer, with a chance of getting $\sim \pi/\sqrt{6}e = 0.472$ closer still by statistical fluctuation.

We find that in higher dimensions (typically greater than ten), the maximum live point is nowhere near the true maximum! However, these shortcomings can be mitigated by using the maximum point in the chain as a *starting point* for a gradient-based optimisation algorithm, which would allow us to leverage the advantages of nested sampling to navigate multimodal landscapes while also rapidly approach the maximum point.

3.3 Current temperature

As previously discussed, nested sampling chooses an annealing schedule that moves at constant thermodynamic speed, which constructs a gradual transition from prior ($\beta = 0$) to posterior ($\beta = 1$). One can expect then that any given point during a run can be associated with a $\beta \in [0, 1]$ that measures how far the current set of samples lies between prior and posterior.

We present here the idea that one can assign each iteration of a nested sampling run i^* with a *current temperature* β^*.

As noted by Skilling, the posterior with respect to $\log X$ $\mathcal{L}^\beta X$ has a peak at $\log X^*$ when

$$\beta^* = -\frac{\mathrm{d} \log X}{\mathrm{d} \log \mathcal{L}} \tag{18}$$

Choice of temperature mapping

The typical behaviour at an intermediate stage of a nested sampling run is that the posterior mass is dominated by a single point. This is because the highest live point is typically several log-likelihoods, and therefore orders of magnitude higher in weight than the next point. Decreasing β, equivalent to increasing the temperature, causes the mass to redistribute more equally across the samples. A natural choice for the present temperature β^* is therefore one where the posterior mass is concentrated around the current dead point. What that constitutes is subjective, and several plausible choices are presented below:

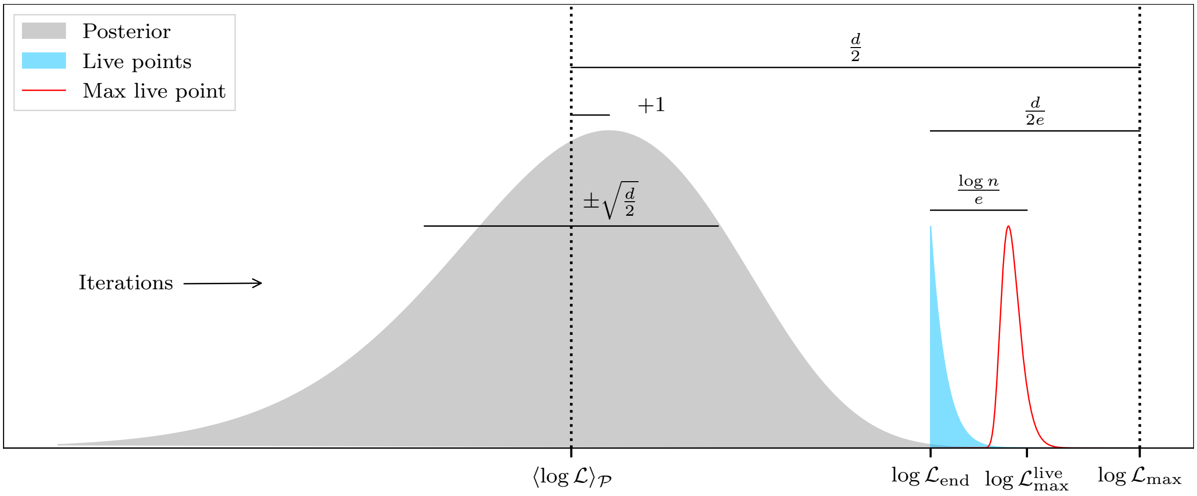


Figure 3. Distribution of samples as a function of $\log \mathcal{L}$, showing the posterior $\mathcal{P}(\log \mathcal{L})$, the distribution of the live points $\pi(\log \mathcal{L} \mid \mathcal{L} > \mathcal{L}_*)$, and the distribution of the maximum likelihood live point $P(\log \mathcal{L}_{\mathrm{max}}^{\mathrm{live}})$

- Setting the posterior averaged log-likelihood $\langle \log \mathcal{L} \rangle_{\mathcal{P}(\beta)}$ to be equal to the current $\log \mathcal{L}^*$.
- Setting the Kullback-Leibler divergence of current set of samples $\mathcal{D}(\beta)$ to be equal to the current $-\log X^*$.
- Setting β so that the set of samples meets some criterion for the remaining posterior mass, for example so that a specified fraction ϵ of the total mass lies beyond the current point.

Note that the second condition is a special case of the third, since $\mathcal{D}_{\mathrm{KL}} = \langle \log \mathcal{L} \rangle_{\mathcal{P}} - \log \mathcal{Z} = -\log X^*$ corresponds to $\mathcal{Z} \sim \mathcal{L}X$ so that $\epsilon \sim 0.5$. One could account for the uncertainty in choice by sampling uniformly between the β^* that correspond to fractions ϵ and $1-\epsilon$ of the posterior mass remaining, since all such choices have a posterior mass

Inferring the current temperature

Transforming $\mathcal{L} \to \mathcal{L}^\beta$ leads to the posterior distribution for $\log X$ to instead follow

$$\mathcal{P}(\log X \mid \beta) = \frac{\mathcal{L}^\beta X}{\int \mathcal{L}^\beta X \, \mathrm{d}\log X}. \tag{19}$$

Using Bayes' rule one can get reverse, which is the probability of the temperature given a set of $\log X$;

$$P(\beta \mid \log X) \propto \mathcal{P}(\log X \mid \beta)\, P(\beta)\,. \tag{20}$$

Evaluating the left hand side for $\log X = \log X^*$ then gives the distribution of β at the current iteration. We note that we would like $P(\beta = 0 \mid \log X^*) = 0$ since the likelihood must always increase, but $\mathcal{L}^\beta X^* = X$

Intermediate sample dimensionalities

One problem with the Bayesian model dimensionality is that it is practically zero for most of the run, only converging at the very end when the posterior mass is no longer concentrated at a single point with zero variance. From the thermodynamic perspective, this is expected; the present number of degrees of freedom only corresponds to the heat capacity at the present, not the final temperature. The number of parameters constrained at an intermediate stage should really be calculated from the *transformed* posterior, not the original set of samples.

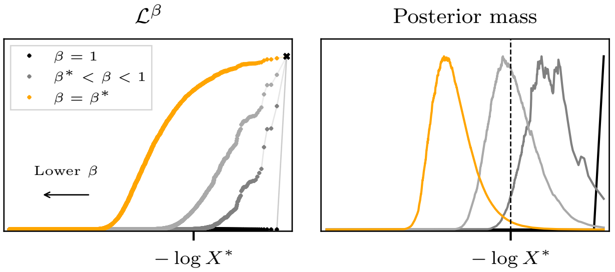


Figure 4. The likelihood profile and posterior mass as β is lowered. The original likelihood is shown in black, which is dominated by the maximum live point. As β decreases, the posterior mass is shifted back from the maximum point to the other points, until so much of the mass has shifted that the present point $\log X^*$ is the termination point. The transformed likelihood and posterior at this point is shown in orange.

Plotted below are the temperature adjusted sample dimensionalities for a spherical Gaussian, for which the ground truth is known.

If this quantity is computed from the untransformed samples, it will output zero since there is effectively only one sample. The transformed posterior, on the other hand, gives a $\tilde{d}$ which is consistent with the distribution of live points, shown in Fig. 5 at various stages for an elongated Gaussian.

Use beta to calculate the current Bayesian model dimensionality, the number of parameters constrained by the current set of samples.

Plot of the model dimensionality throughout the run for an elongated Gaussian, comparing to that obtained with beta = 1. Find that it matches up well with the apparent dimension of the current set of samples.

Find that a feature of nested sampling is that parameters with high variance are initially 'hidden'. Compression occurs in the direction which is most likely to have a sample of higher likelihood. Much easier to find a better point along the direction of a parameter that is initially poorly constrained. So compressions do not occur in parameters that are already well-constrained; appears as though they are uniformly distributed.

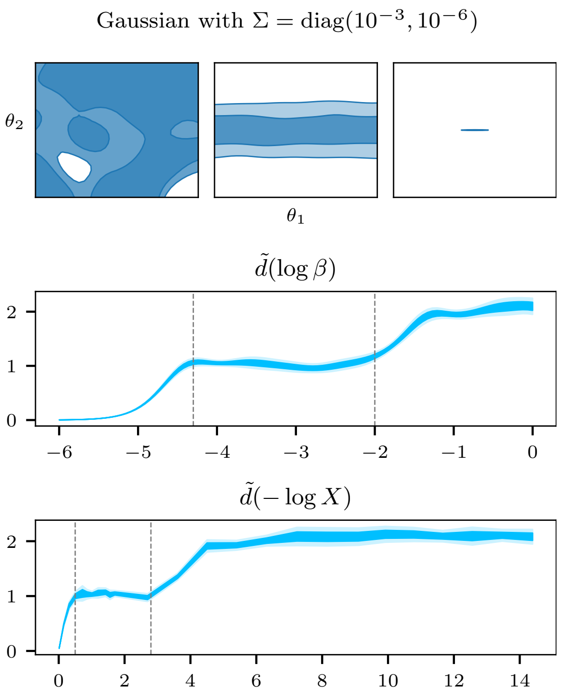


Figure 5. The Bayesian model dimensionality $\tilde{d}$ as a function of β for a Gaussian two parameters of variance 10^{-3} and 10^{-6}. The number of contrained samples can be seen to increase from zero to the true dimensionality of 2, as the run progresses.

4 METHOD

The analysis in Section 3.1 shows that our knowledge of the posterior at an intermediate iteration is limited by the maximum log-likelihood live point, which until just before the end is quite far from the posterior bulk. However, we can get an approximate image of the full posterior ahead of time, simply by *extrapolating* the known likelihood profile; that is, the trajectory of $\mathcal{L}(X)$ traced out by the live and dead points.

One would never use this approximate posterior to do inference, since the extrapolated profile is much less exact than if one were to simply finish the run. However, it is more than sufficient for making a prediction for roughly how much evidence remains, and hence the point at which the run should terminate. Quantitatively, this proceeds as follows: we fit a model function $f(X, \boldsymbol{\theta})$ with some parameters $\boldsymbol{\theta}$ to the known likelihood profile, which allows us to express the prior volume we need to compress to as

$$\Delta\mathcal{Z} = \epsilon\mathcal{Z}_{\mathrm{tot}}, \tag{21}$$

$$\int_0^{X_\mathrm{f}} f(X,\boldsymbol{\theta})\,\mathrm{d}X = \epsilon\left(\int_0^{X_i} f(X,\boldsymbol{\theta})\,\mathrm{d}X + \mathcal{Z}_{\mathrm{dead}}\right), \tag{22}$$

where X_i is the volume of the iteration we have currently compressed to, and $\mathcal{Z}_{\mathrm{dead}}$ 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.

Once X_f is known, the corresponding iteration count depend on the live point schedule. The conversion is easiest in the constant n_{live} case; at each iteration $\log X$ decreases by $1/n_{\mathrm{live}}$, so the total number of iterations N_f will be

$$N_\mathrm{f} = -n_{\mathrm{live}}\log X_\mathrm{f}. \tag{23}$$

4.1 Choosing the extrapolation

One has freedom to choose the form of the fitting function $f(X, \boldsymbol{\theta})$. Following the previous analysis, we use a Gaussian profile in the form of Eq. (A1). It has the advantage that one can speed up its regression analytically, as well as the fact that many real likelihoods are Gaussian or near-Gaussian. We allow d to freely vary to increase flexibility, and reflect the fact that while the dimensionality of a likelihood is fixed, the equivalent Gaussian dimensionality may differ (Handley & Lemos 2019). Regardless of motivation, the proof is in the pudding, so we shall justify the choice on the accuracy of predictions in Section 5.

We must also choose the set of points over which we fit. Using all of the points we have is a poor choice, because unless the profile is actually Gaussian the extrapolation will be insensitive to changes in the profile. A natural choice is then to use the live points only, since these represent the most up-to-date likelihood profile. For a given set of live points, we have a set of live likelihoods $\{\log\mathcal{L}_i\}$, as well as a set of $\{X_i\}$ drawn from the distribution (4). The parameters of the fitting function can be found by minimising the cost function

$$C^2(\boldsymbol{\theta}) = \sum_i \left|\log\mathcal{L}_i - \log f(X_i,\boldsymbol{\theta})\right|^2 \tag{24}$$

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 σ in terms of the optimal values of d:

$$\sigma^2 = \frac{N\sum_i X_i^{4/d} - \left(\sum_i X_i^{2/d}\right)^2}{2\sum_i \log\mathcal{L}_i \sum_i X_i^{2/d} - 2N\sum_i X_i^{2/d}\log\mathcal{L}_i}, \tag{25}$$

$$\log\mathcal{L}_{\max} = \frac{1}{N}\sum_i \log\mathcal{L}_i + \frac{1}{2N\sigma^2}\sum_i X_i^{2/d}, \tag{26}$$

(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.

4.2 The termination prior volume

Once we have obtained the best-fit $\boldsymbol{\theta}$, the endpoint can be calculated in the manner described in Eq. (22);

$$\epsilon = \frac{\int_0^{X_\mathrm{f}} \mathcal{L}_{\max}\exp\left(-X^{2/d}/2\sigma^2\right)\,\mathrm{d}X}{\int_0^{X_i} \mathcal{L}_{\max}\exp\left(-X^{2/d}/2\sigma^2\right)\,\mathrm{d}X + \mathcal{Z}_{\mathrm{dead}}}. \tag{27}$$

The integrals have the analytic solution

$$\int_0^{X_k} \mathcal{L}_{\max}\exp\left(-X^{2/d}/2\sigma^2\right)\,\mathrm{d}X = \frac{d}{2}\cdot\left(\sqrt{2}\sigma\right)^d\cdot\gamma_k \tag{28}$$

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 γ and a few more steps of algebra, we arrive at

$$\log X_\mathrm{f} = \frac{d}{2}\log 2\sigma^2 + \log\gamma^{-1}\left(\frac{d}{2},\ \epsilon\gamma_i + \frac{\epsilon\mathcal{Z}_{\mathrm{dead}}}{\left(2\sigma^2\right)^{d/2}\mathcal{L}_{\max}}\right), \tag{29}$$

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

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

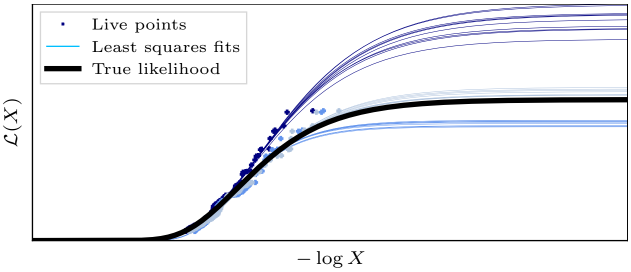


Figure 6. Sources of the prediction uncertainty. The navy points are the live points, plotted for three different instances of $\{X_i\}$ drawn from its distribution. The light blue lines show the uncertainty in the least squares fits over each instance, which is much less than the uncertainty in the spread of the X samples due to long term drift.

4.3 Prediction uncertainties

We saw in section X that our knowledge of the posterior is limited by the live point of minimum volume, which is only $O(\log n_{\mathrm{live}})$ less than the log-compression of the dead point. Endpoint prediction as we have just done relies on our hope that the likelihood will continue to behave as it has done up to this point, but this may well not be the case; a sharp phase change, for instance, will be our undoing.

We will therefore restrict ourselves to quantifying the uncertainties associated with how we have *modelled* the known likelihood profile. These have two sources: the variance in the prior volume estimates, and the uncertainty in the least squares regression. The former is by far the larger source of error, since the uncertainty in nested sampling is dominated not by the scatter of the individual points but by long-term drift (as shown in Fig. 6).

To get the uncertainty, we can therefore ignore the least squares uncertainty and simply sample from $\Pr(\boldsymbol{X})$, finding the best-fit $\boldsymbol{\theta}$ and endpoint for each sample, then using this set of estimates to get a mean and variance in the endpoint estimate.

5 RESULTS

Now that we have a method for predicting the endpoint, we can test it on a range of likelihoods. We shall first present results for cosmological chains in Section 5.2, for which the method works well. Section 5.3 will then show some toy examples that are designed to reveal some of the limitations of the method.

5.1 Gaussians

We will first test our prediction model on truly Gaussian likelihoods to show that it is able to recover the true parameters as well as the correct endpoint. Let us begin with a spherically symmetrical Gaussian which matches the form of the fitting function, which we can use to check that the correct parameters are inferred when the model is exact. We fix the parameters $\log \mathcal{L}_{\max} = 0$ and $\sigma = 0.01$ and vary the dimension to see the impact on the predictions.

The inferred parameters are shown in Fig. 7. The uncertainty in d and σ remain large throughout the run, but the endpoint predictions converge because we have already accumulated most of the evidence.

Predictions are shown in Fig. 11a, where the x-axis plots the negative log-compression as a measure of the progress, so that the run proceeds in the rightward direction. The y-axis shows the predicted log-compression $\log \hat{X}_{\mathrm{f}}$ at which the termination condition is met.

The true endpoint is recovered to within 2σ for all runs, but the uncertainty is greater for the higher dimension and $\mathcal{D}_{\mathrm{KL}}$ likelihoods. This reflects the fact that the live points typically probe only $\sim \log n_{\mathrm{live}}$ away from the current dead point. For larger $\mathcal{D}_{\mathrm{KL}}$ therefore, the set of samples at any given iteration before the end are quite far from the posterior bulk, and hence poorly resemble the shape of the true posterior.

5.1.1 *Elongated Gaussian*

Consider same example as that presented in **??**. The predictions are initially awry because samples from three of the parameters are uniform. Cannot determine whether these parameters are unconstrained or simply high variance at the midway point. So the predictions are fundamentally limited by the information contained in the samples, not by the prediction method.

Comparison to extrapolating evidence increments

Seasoned users of nested sampling might be curious how the method compares to simply extrapolating the increments of evidence to roughly estimate when the evidence converges. We do this for a spherical Gaussian as a comparison to our method. At an intermediate stage of the run, the most recent outputs might look something like that shown in the first two columns of the table in Fig. 8. Extrapolating those data to a linear and exponential profile yields endpoint estimates plotted in the graph to the right.

The linear extrapolation is clearly an underestimate, since it fails to account for the long tail of the nonlinear profile. The increments are also not quite exponential, since the exponential fit leads to a large overprediction. The predicted endpoint over the course of a run for $d = 16$, $\sigma = 0.01$, as shown in Fig. 9, shows the same result. One might expect an average to be more accurate, but this tends to be biased towards the exponential prediction, and there is no obvious choice of weighting that would fix this.

More importantly, we find that for real likelihoods which have an element of noise the extrapolation often diverges, for instance when the increments do not monotonically decrease. Directly extrapolating the evidence increments is therefore far less stable than the previous method, and generally not a reliable method for prediction.

5.2 Cosmological examples

Having established that the prediction framework works for Gaussian likelihoods, we now turn to some cosmological examples, namely chains based on the Planck 2018 dataset as described in Handley (2021). Predictions are shown in Fig. 11b, with the correct endpoint inferred to within error before the halfway point for all likelihoods.

Most predictions fail to converge at early stages of the run. We find that this is not a consequence of the prediction model, but due to the effective dimensionality of the samples being lower than the true value at the beginning of the run, as was seen in Section 3.3.

Compressive phase transitions

The initial dimensionality of the $\mathcal{L}(X)$ curve is so extrapolating from the curve would fail to account for the true dimensionality of the likelihood, giving inaccurate predictions.

The extrapolation now accounts for the true dimensionality, so the

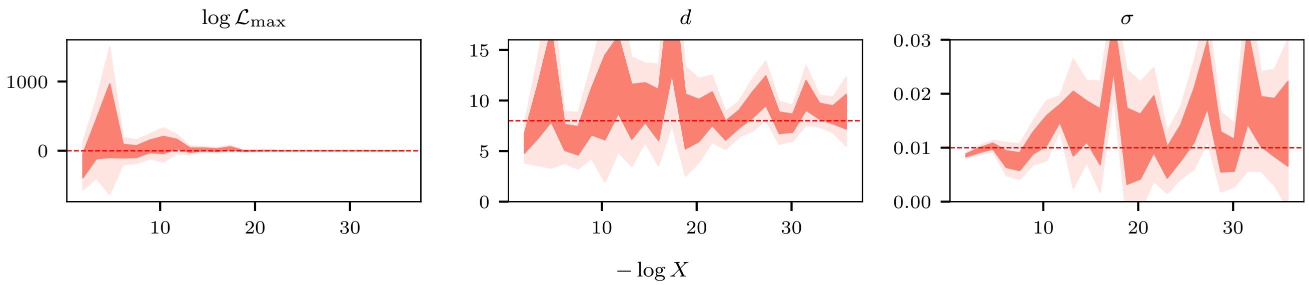


Figure 7. Parameter inferences for the Gaussian fitting function in the $d = 8$ case. The true parameters fall within the standard error bars for practically every point, as seen in comparison to the true parameter values shown in the dotted red. The uncertainties remain relatively constant for d and σ, but the endpoint predictions converge because we have already accumulated most of the evidence.

iteration	log Z	d log Z
5000	-1435.8	190.8
5500	-1264.6	171.2
6000	-1123.7	140.9
6500	-991.5	132.2
7000	-885.0	106.6
7500	-790.3	94.7
8000	-702.6	87.7
8500	-619.7	82.9
9000	-551.8	67.9
9500	-492.7	59.1

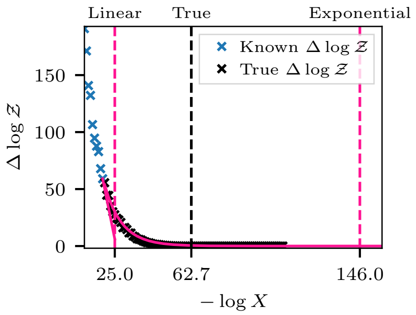


Figure 8. Extrapolating the increments of evidence. The left column shows the output of a nested sampling run, and the right column shows the extrapolation

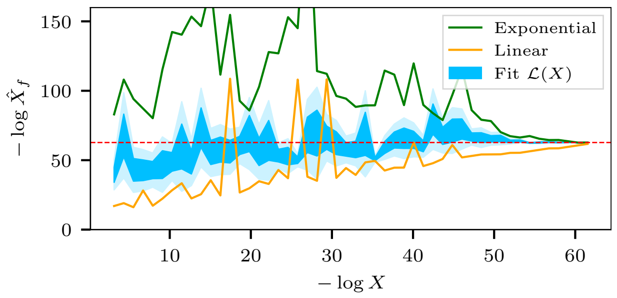


Figure 9. Extrapolating the increments of evidence. The left column shows the output of a nested sampling run, and the right column shows the extrapolation

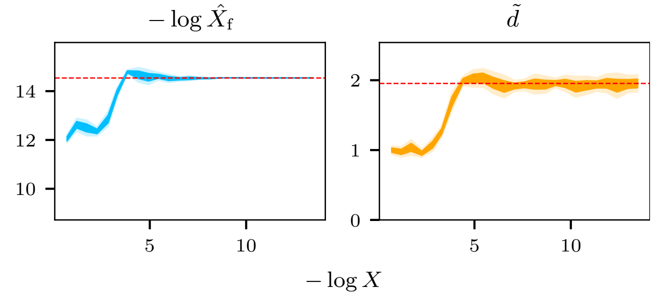


Figure 10. Endpoint predictions and effective dimensionality as a function of compression for a two-dimensional Gaussian in a box prior $\in [0, 1]$. The parameter variances are 10^{-2} and 10^{-5} respectively, leading to compression only along one direction initially, which is reflected in $\tilde{d}$. The endpoint only converges when the dimensionality does; before that point the samples genuinely suggest that the likelihood is one-dimensional.

predictions become accurate. The trouble is that before this has happened, the samples really do behave as if they are only constrained in a single dimension! It could really be the case that there is no second parameter, and without our extra knowledge there is no way to detect whether there are two, or even three or more unseen parameters hiding at higher compressions

The different rates of compression for different parameters therefore induce an effective phase transition in the number of degrees of freedom. One can put this on a more quantitative footing by considering the *effective dimensionality* of the posterior. Define the Bayesian model dimensionality $\tilde{d}$ as the posterior variance of the information content, which is given by

Fig. 10 plots the effective dimensionality alongside the endpoint predictions, where it is evident that the predictions are directly dependent on $\tilde{d}$. Introducing correlations between the parameters (not shown) causes the transition to be less sharp, since the more constrained parameter is able to 'drag' the other parameter with it, allowing the predictions can be converge shortly before the dimensionality.

5.3 Edge cases

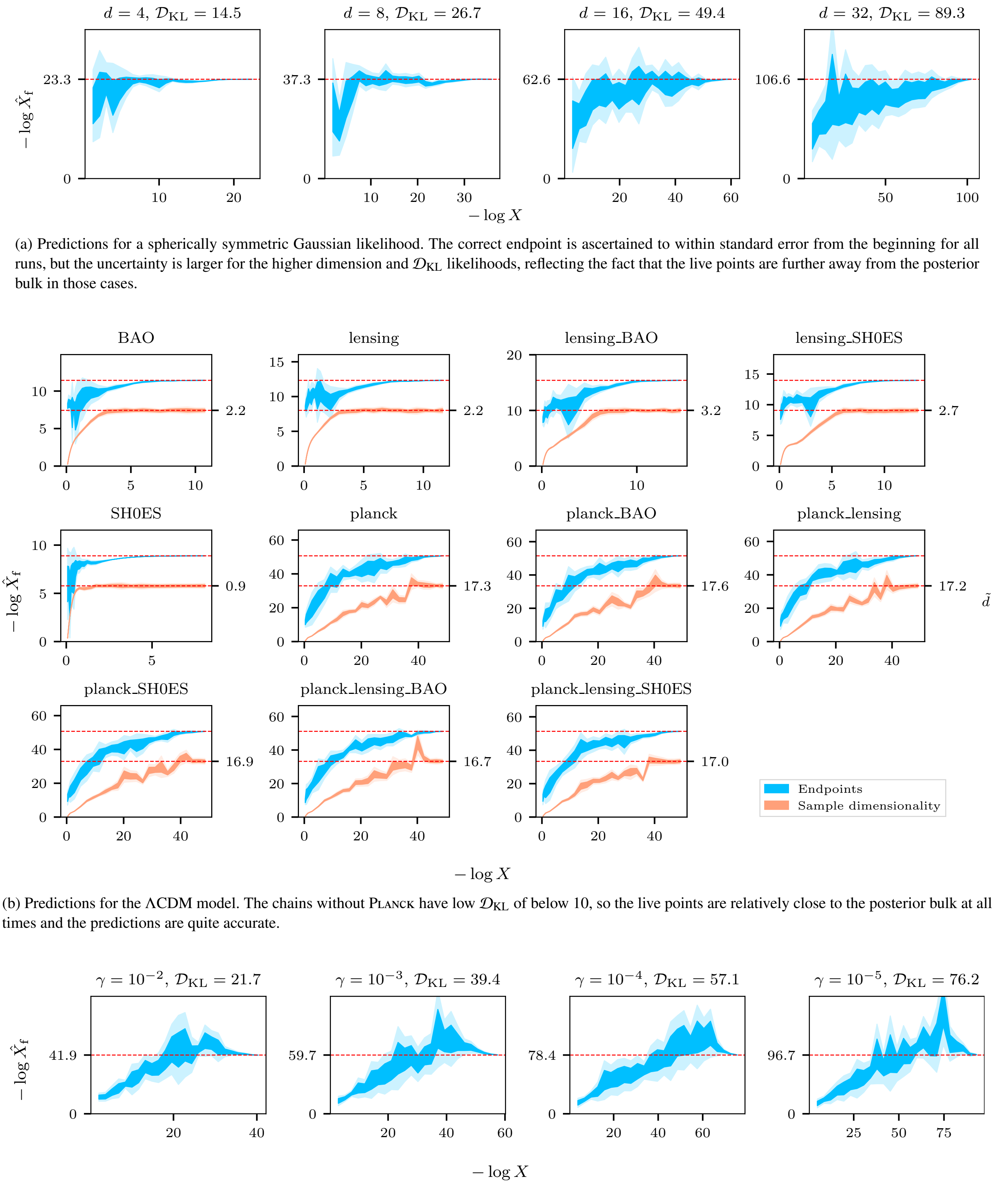
In the cosmological cases, underestimated predictions at early stages was due to the difference in parameter compression rates. However, there are also cases where the form of the fitting function causes problems.

One such instance is a spherical Cauchy likelihood. In fact, at early stages direct least squares fitting fails completely by predicting a remaining evidence that overflows. This can be explained by the fact that a long tail of a Cauchy resembles a Gaussian of much higher dimension, and inferring the dimensionality is unstable. Since we have a stable measure of dimension from the Bayesian model dimensionality, a partial fix is to set $d = \tilde{d}$ instead of inferring it, then using Eqs. (25) and (26) to get $\log \mathcal{L}_{\max}$ and σ.

Results are shown in Fig. 11c, where the likelihood has the form

$$\log \mathcal{L} = \log \mathcal{L}_{\max} - \frac{1+d}{2} \log\left(1 + \frac{X^2}{\gamma^2}\right), \tag{30}$$

choosing $d = 10$ and allowing γ to vary. However, the extrapolations still fall outside fall outside the standard error range despite isotropic compression. The issue in this case is that the Gaussian fitting function is unable to capture the true shape of the likelihood. A more comprehensive fix would be to use a more flexible fitting function that is able to capture a broader family of likelihoods, which is outside the scope of this paper. Nevertheless, the order of magnitude is still accurate, which remains highly useful to the end user.



(a) Predictions for a spherically symmetric Gaussian likelihood. The correct endpoint is ascertained to within standard error from the beginning for all runs, but the uncertainty is larger for the higher dimension and $\mathcal{D}_{\mathrm{KL}}$ likelihoods, reflecting the fact that the live points are further away from the posterior bulk in those cases.

(b) Predictions for the ΛCDM model. The chains without Planck have low $\mathcal{D}_{\mathrm{KL}}$ of below 10, so the live points are relatively close to the posterior bulk at all times and the predictions are quite accurate.

(c) Predictions for a Cauchy likelihood. The parameters of the fitting function are obtained by fixing $d = \tilde{d}$, then least squares optimising $\log \mathcal{L}_{\max}$ and σ.

Figure 11. Endpoint predictions for different likelihoods. The x-axis shows the negative log-compression, and the y-axis the predicted endpoint $\log \hat{X}_{\mathrm{f}}$. The shaded regions show the 1 and 2σ uncertainties, while the red dashed line shows the true value.

6 CONCLUSIONS

We have presented a method for predicting the endpoint of a nested sampling run, based on extrapolating the known likelihood profile into higher likelihood regions using a Gaussian fitting function.

The method in general converges to the correct endpoint by about the halfway point, and gets the correct order of magnitude throughout. Predictions are poor when the effective dimension of the samples is lower than the true posterior, but this is a limitation of the information content in the samples rather than our method.

Further work can be done to incorporate more flexible fitting functions to extrapolate non-Gaussian likelihoods, as well as developing a more rigorous framework to choose between the inferred d and $\tilde{d}$.

REFERENCES

Akrami Y., Scott P., Edsjö J., Conrad J., Bergström L., 2010, Journal of High Energy Physics, 2010

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

Feroz F., Cranmer K., Hobson M., de Austri R. R., Trotta R., 2011, Journal of High Energy Physics, 2011

Handley W., 2019, Journal of Open Source Software, 4, 1414

Handley W., 2021, Physical Review D, 103

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., 2018, Monthly Notices of the Royal Astronomical Society

Petrosyan A., Handley W., 2022, Physical Sciences Forum, 5

Pártay L. B., Bartók A. P., Csányi G., 2010, Efficient sampling of atomic configurational spaces (arXiv:0906.3544)

Skilling J., 2006, Bayesian Analysis, 1

APPENDIX A: DERIVATION OF THE TERMINATION LOG-LIKELIHOOD

Assume a Gaussian likelihood

$$\log\mathcal{L} = \log\mathcal{L}_{\max} - \frac{X^{2/d}}{2\sigma^2}. \tag{A1}$$

The distribution of the true posterior in $\log\mathcal{L}$ is

$$P(\log\mathcal{L}) = \frac{1}{\Gamma(\frac{d}{2})} e^{\log\mathcal{L}-\log\mathcal{L}_{\max}} (\log\mathcal{L}_{\max} - \log\mathcal{L})^{\frac{d}{2}-1} \tag{A2}$$

i.e. $2(\log\mathcal{L}_{\max} - \log\mathcal{L}) \sim \chi^2_d$, which is the distribution of the weighted samples. The posterior average and variance of $\log\mathcal{L}$ are given by

$$\langle\log\mathcal{L}\rangle_{\mathcal{P}} = \log\mathcal{L}_{\max} - \frac{d}{2}, \quad \mathrm{Var}(\log\mathcal{L})_{\mathcal{P}} = \frac{d}{2}. \tag{A3}$$

Meanwhile, the live points are uniformly distributed over the constrained prior and hence have probability distribution

$$P(\log\mathcal{L}) = \frac{d}{2}\frac{(\log\mathcal{L}_{\max} - \log\mathcal{L})^{\frac{d}{2}-1}}{(\log\mathcal{L}_{\max} - \log\mathcal{L}_*)^{\frac{d}{2}}} \quad [\log\mathcal{L}_* < \log\mathcal{L} < \log\mathcal{L}_{\max}], \tag{A4}$$

It is helpful at this stage to define a parameter

$$y = \frac{\log\mathcal{L} - \log\mathcal{L}_*}{\log\mathcal{L}_{\max} - \log\mathcal{L}_*} \tag{A5}$$

as a normalised measure of how far a point is between the latest dead point and the maximum log-likelihood, with $y = 0$ corresponding to $\mathcal{L}_*$ and $y = 1$ to $\mathcal{L}_{\max}$, so that

$$P(y) = \frac{d}{2}(1-y)^{\frac{d}{2}-1} \quad [0 < y < 1]. \tag{A6}$$

We now seek the distribution for the maximum likelihood of the live points, $\log\mathcal{L}^{\mathrm{live}}_{\max}$. Using the result that the maximum of n variables with cumulative distribution $F(y)$ follows $\frac{d}{dy}(1-(1-F(y))^n)$, we obtain

$$P(y^{\mathrm{live}}_{\max}) = \frac{nd}{2}(1-y^{\mathrm{live}}_{\max})^{\frac{d}{2}-1}\left(1-(1-y^{\mathrm{live}}_{\max})^{\frac{d}{2}}\right)^{n-1} \quad [0 < y^{\mathrm{live}}_{\max} < 1], \tag{A7}$$

which in the limit of large live points and dimensions can be roughly summarised by

$$\lim_{d,n\to\infty} y^{\mathrm{live}}_{\max} \sim \frac{2\log n}{d} \pm \sqrt{\frac{2}{3}\frac{\pi}{d}}. \tag{A8}$$

This shows that in general the live points are nowhere near the maximum log-likelihood at any iteration, though they do steadily squeeze the interval $[\log\mathcal{L}_*, \log\mathcal{L}_{\max}]$. In particular, in high dimensions n only gets us harmonically/logarithmically closer, whilst d pushes us linearly further away.

This remains true even at the end of the nested sampling run. To see this, we write the halting condition as:

$$f = \frac{\int_0^{X_{\mathrm{end}}} \mathcal{L}\,\mathrm{d}X}{\int_0^{\infty} \mathcal{L}\,\mathrm{d}X}. \tag{A9}$$

Note that we have assumed that prior effects are negligible (so $1 = \infty$), and that $f \ll 1$ so that the denominator is approximately the accumulated evidence. Computing this for Eq. (A1) we find the answer in terms of lower incomplete gamma functions

$$f = 1 - \frac{\Gamma_{d/2}\left(\frac{X_{\mathrm{end}}^{2/d}}{2\sigma^2}\right)}{\Gamma\left(\frac{d}{2}\right)}. \tag{A10}$$

Taking the $X_{\mathrm{end}} \ll (\sqrt{2}\sigma)^d$ limit (almost certainly valid at termination) we find

$$\lim_{X_{\mathrm{end}}\ll(\sqrt{2}\sigma)^d} f \approx \frac{X_{\mathrm{end}}}{(\sqrt{2}\sigma)^d\,\Gamma(1+\frac{d}{2})} = \frac{(\log\mathcal{L}_{\max} - \log\mathcal{L}_{\mathrm{end}})^{\frac{d}{2}}}{\Gamma(1+\frac{d}{2})}. \tag{A11}$$

We thus have an expression relating $\mathcal{L}_{\mathrm{end}}$ at termination to the termination fraction f. This becomes yet more pleasing in the large d limit, since $f^{2/d} \to 1$, we find via a Stirling approximation:

$$\lim_{d\to\infty} \log\mathcal{L}_{\mathrm{end}} \approx \log\mathcal{L}_{\max} - \frac{d}{2e}. \tag{A12}$$

In the event that we keep f in, we replace $\frac{d}{2e} \to \frac{d}{2e}f^{2/d}$, so we can of course battle the $\frac{d}{2e}$ term, but this becomes exponentially difficult in high dimensions.

Putting this together, taking $\mathcal{L}_*$ in Eq. (A5) to be $\mathcal{L}_{\mathrm{end}}$, and combining this with Eq. (A8) we find

$$\boxed{\log\mathcal{L}^{\mathrm{live}}_{\max} \approx \log\mathcal{L}_{\max} - \frac{d}{2e} + \frac{\log n}{e} \pm \frac{\pi}{\sqrt{6e}}}, \tag{A13}$$

showing that in general nested sampling will finish at a contour $d/2e$ away from the maximum log-likelihood, and the final set of n live points can get you $\log(n)/2e$ closer, with a chance of getting $\sim \pi/\sqrt{6e} = 0.472$ closer still by statistical fluctuation. This general anatomy is shown in Fig. 3.