Approximating the end of nested sampling

Zixiao Hu, Artyom Baryshnikov, Will Handley

4 October 2023

ABSTRACT

Key words: methods: data analysis – methods: statistical

```
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.

1 INTRODUCTION

Nested sampling is a multi-purpose algorithm invented by John Skilling which simultaneously functions as a probabilistic sampler, integrator and optimiser (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 by *counting*, which makes high-dimensional integration feasible. It also avoids problems faced by traditional Bayesian algorithms, such as multi-modality.

The order of magnitude runtime of an algorithm, that is, whether termination is hours or weeks and months away, is of high importance to the end user. Currently, existing implementations of nested sampling (Feroz et al. 2009; Handley et al. 2015) either do not give an indication of runtime, or only provide crude measures of progress that do not directly correspond to the runtime.

This paper sets out a more principled manner of endpoint estimation for nested sampling at each intermediate stage, the key idea being to use the existing samples to predict the likelihood in the region we have yet to sample from. Outline of paper.

2 BACKGROUND

Let us begin with a brief description of the nested sampling algorithm to establish the necessary notation. 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 $n_{\rm live}$ *live points* sampled from the prior, which can vary in number throughout the run (Higson et al. 2019). 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

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

where $X(\mathcal{L})$ is the fraction of the prior with a likelihood 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 $n_{\rm live}$ points drawn from $[0, 1]$, so follow the distribution

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

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 $\boldsymbol{t}$. For any given problem, the uncertainty in $\log \mathcal{Z}$ is proportional to $1/\sqrt{n_{\rm live}}$, so $n_{\rm live}$ sets the resolution.

3 THE ANATOMY OF A NESTED SAMPLING RUN

The following section aims to make an inventory of the information available to us at an intermediate iteration i^*, which we shall eventually use to make endpoint predictions. We present an anatomy of the progression of a nested sampling run in terms of the prior volume compression, the log-likelihood increase, the inferred temperature schedule, and the dimensionality of the samples.

3.1 Prior volume

The key feature of nested sampling is that the sampling is controlled by prior volume compression. The task is to find the posterior typically lying in a tiny fraction of the prior volume, a total compression which is quantified by the average information gain, or *Kullback-Leibler divergence*:

$$\mathcal{D}_{\rm KL} = \int \mathcal{P}(\theta) \log \frac{\mathcal{P}(\theta)}{\pi(\theta)} \, \mathrm{d}\theta. \tag{6}$$

The bulk of the posterior lies within a prior volume $X = e^{-\mathcal{D}_{\rm KL}}$, which is the target compression. One gets there by iteratively taking steps of size $\Delta \log X_i = -1/n_i$, so that

$$\mathrm{E}(\log X_i) = -\sum_{k=1}^{i} \frac{1}{n_k}, \quad \mathrm{Var}(\log X_i) = \sum_{k=1}^{i} \frac{1}{n_k^2}. \tag{7}$$

A constant step size in $\log X$ corresponds to a geometrically decreasing measure for the dead points (as shown in fig X), which is exactly needed to overcome the curse of dimensionality.

Dead measure plot

The same is not true for the live points, which are uniformly distributed in prior volume. As a result, the maximum live point is found at

$$\mathrm{E}[\log X_{\rm min}^{\rm live}] = \mathrm{E}[\log X_*] - \sum_{k=1}^{n_{\rm live}} \frac{1}{n_k} \approx -\frac{i_*}{n_{\rm live}} - \log n_{\rm live} - \gamma, \tag{8}$$

with variance

$$\mathrm{Var}[\log X_{\rm min}^{\rm live}] = \mathrm{Var}[\log X_*] + \sum_{k=1}^{n_{\rm live}} \frac{1}{n_k^2} \approx \frac{i_*}{n_{\rm live}^2} + \frac{\pi^2}{6}, \tag{9}$$

where the large $n_{\rm live}$ limit is taken for the approximation to the harmonic series, γ being the Euler-Mascheroni constant.

Hence the live points only get us a factor of $\log n_{\rm live}$ closer to the posterior bulk. In other words, it is not until we are around $\log n_{\rm live}$ away from $\log X = \mathcal{D}_{\rm KL}$ that the samples look anything like the posterior. One can see from Eq. (6) that the divergence increases linearly with dimension, so for large dimensionalities and typical live point numbers $\lesssim 1000$, this does not happen until near the end of the run.

Intuitively, it 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 Log-likelihood

It is also useful to observe the distribution of the live and dead points in log-likelihood. We choose to examine a representative case of the d-dimensional multivariate Gaussian:

$$\log \mathcal{L} = \log \mathcal{L}_{\rm max} - X^{2/d}/2\sigma^2 \tag{10}$$

with maximum point $\log \mathcal{L}_{\rm max}$ and lengthscale σ to get an insight into the analytics. The likelihood monotonically increases towards $\log \mathcal{L}_{\rm max}$, with the posterior samples eventually concentrated around the bulk at

$$\langle \log \mathcal{L} \rangle_{\mathcal{P}} = \log \mathcal{L}_{\rm max} - \frac{d}{2}, \quad \mathrm{Var}(\log \mathcal{L})_{\mathcal{P}} = \frac{d}{2}. \tag{11}$$

We can again find the size of each step in log-likelihood, as well as the expected location of the maximum live point. Define a likelihood normalised by the distance to the maximum

$$y = \frac{\log \mathcal{L} - \log \mathcal{L}_*}{\log \mathcal{L}_{\rm max} - \log \mathcal{L}_*} \tag{12}$$

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 $\log \mathcal{L}_{\rm max}$. At each iteration, y increases by roughly

$$\lim_{n_{\rm live}\to\infty} \Delta y \approx \frac{\mathrm{d}y}{\mathrm{d}\log X} \Delta \log X = \frac{2}{d n_{\rm live}} \tag{13}$$

in the large $n_{\rm live}$ limit. We can again sum the harmonic series to see that the maximum live point is expected to be at $y = 2\log n_{\rm live}/d$: a very small fraction in high dimensions, showing again that until the end the live points are far from the posterior bulk, and certainly nowhere near the maximum.

The above also implies that the normalised distance between the highest and second highest live point is roughly $2/d$. Before reaching the posterior bulk, $\log \mathcal{L}_{\rm max} - \log \mathcal{L}^* > d/2$, so the log-likelihood distance between the two points is larger than one, hence at least an order of magnitude apart in likelihood. It is therefore typically the case that nearly all of the posterior mass is concentrated in a single point, the maximum live point, until the very end of the run when the prior volumes have shrunk enough to compensate.

Aside: nested sampling as a maximiser

Previous literature (Akrami et al. 2010; Feroz et al. 2011) has explored the potential for nested sampling to be used as a global maximiser, given its ability to handle multi-modalities. In particular, the latter authors emphasised that posterior samplers such as nested sampling find the bulk of the *mass*, not the maximum of the distribution, but that this can be remedied by tightening the termination criterion. We now use the machinery we have developed to put this statement in more quantitative terms.

A more rigorous derivation (see Appendix) shows that the maximum live point has mean and variance

$$\lim_{d, n_{\rm live}\to\infty} y_{\rm max}^{\rm live} \sim \frac{2\log n_{\rm live}}{d} \pm \sqrt{\frac{2}{3}}\frac{\pi}{d}. \tag{14}$$

Now let us take the dead point to be the termination point with likelihood $\log \mathcal{L}_{\rm end}$ and prior volume $X_{\rm end}$, so that

$$\epsilon = \frac{\int_0^{X_{\rm end}} \mathcal{L} \, \mathrm{d}X}{\int_0^{\infty} \mathcal{L} \, \mathrm{d}X}. \tag{15}$$

Note that we have assumed that prior effects are negligible (so $1 = \infty$), and that $\epsilon \ll 1$ so that the denominator is approximately the accumulated evidence. Computing this for Eq. (10), we find the answer in terms of lower incomplete gamma functions

$$\epsilon = 1 - \frac{\Gamma_{d/2}\left(X_{\rm end}^{2/d}/2\sigma^2\right)}{\Gamma(d/2)}. \tag{16}$$

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

$$\lim_{X_{\rm end}\ll(\sqrt{2}\sigma)^d} \epsilon \approx \frac{X_{\rm end}}{(\sqrt{2}\sigma)^d\,\Gamma(1+\frac{d}{2})} = \frac{(\log\mathcal{L}_{\rm max} - \log\mathcal{L}_{\rm end})^{\frac{d}{2}}}{\Gamma(1+\frac{d}{2})}. \quad (17)$$

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

$$\lim_{d\to\infty} \log\mathcal{L}_{\rm end} \approx \log\mathcal{L}_{\rm max} - \frac{d}{2e}. \quad (18)$$

In the event that we keep ϵ in, we replace $\frac{d}{2e} \to \frac{d}{2e}\epsilon^{2/d}$, 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. (12) to be $\mathcal{L}_{\rm end}$, and combining this with Eq. (14) we find

$$\boxed{\log\mathcal{L}^{\rm live}_{\rm max} \approx \log\mathcal{L}_{\rm max} - \frac{d}{2e} + \frac{\log n}{e} \pm \frac{\pi}{\sqrt{6}e}}, \quad (19)$$

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

3.3 Temperature

Motivations

As shown in the previous section, midway through the run nearly all of the posterior mass is concentrated at a single point. However, this does not capture the *structure* of the posterior that has been explored and all of the information it provides.

We have the potential to fix this because nested sampling is invariant to monotonic transformations; softening the likelihood via $L \to L^\beta$ brings the likelihoods closer together. Large β and small β both discard information, but there is an inbetween range of β which is significant, because it causes the posterior mass to be concentrated at the current point; see plot.

At this point relevant to note the correspondence between Bayesian inference and statistical mechanics, from which the above transform is derived. If one takes $-\log\mathcal{L}$ to be the energy E_i etc, then β is the inverse temperature $1/T$. Softening or hardening is equivalent to setting the temperature of the samples, which one can easily do in NS because it is athermal; get the posterior for $\mathcal{L}^\beta$ trivially by reweighting.

Bayesian inference has a correspondence to statistical mechanics in that one can equate the parameters to microstates i, the negative log-lihood to the microstate energy E_i, and the prior to the density of states g_i. One can generalise Bayes' theorem to

$$p(E_i) = \frac{g_i e^{-\beta E_i}}{Z(\beta)} \quad \leftrightarrow \quad \mathcal{P}_\beta(\theta) = \frac{\mathcal{L}^\beta(\theta)\pi(\theta)}{Z(\beta)} \quad (20)$$

Clear similarity to thermodynamic integration, which uses temperature as a control parameter to construct intermediate distributions between the prior at $\beta = 0$ and posterior at $\beta = 1$. NS uses compression instead of temp, sidestepping difficulty of designing the annealing schedule. However, it still does something similar, starting with unweighted prior samples and ending with likelihood-weighted prior samples i.e. the posterior, during which β presumably increases from 0 to 1. Inferring the temperature from above, one would hope to get another quantity that determines the progress of a run.

Important difference to annealing is that here we are working backwards; instead of choosing a temperature to sample at, we obtain the temperature from the samples after the fact. An useful discussion can be found in Habeck (2015), which we summarise here as a starting point. In contrast to annealing methods which attempt to sample from a series of canonical ensembles $\pi(\theta)\exp -\beta\mathrm{E}(\theta)$, nested sampling evolves a series of *microcanonical* ensembles and computing the density of states by tracking the compression.

We now present several methods of obtaining the temperature that one can plausibly consider to be the current temperature of a nested sampling run.

A. *Microcanonical temperature*

As discussed in Skilling's original paper, the density of states, given by the negative slope of the log-likelihood curve, directly corresponds to a temperature;

$$\beta^* = -\left.\frac{\mathrm{d}\log X}{\mathrm{d}\log\mathcal{L}}\right|_{\log\mathcal{L}^*}. \quad (21)$$

This is the microcanonical temperature $\partial S/\partial E$, where the volume entropy $S = \log X$ and the energy $E = -\log\mathcal{L}$. Its value can be easily obtained via finite difference of the $\log\mathcal{L}$ and $\log X$ intervals, albeit subject to an arbitrary window size for the differencing (in practice, we find that above a threshold of around 10 iterations the result is fairly insensitive to this choice).

B. *Canonical temperature*

Briefly considered by Habeck, it is the temperature of an ensemble whose average energy is the current energy, which one can obtain by inverting

$$\langle\log\mathcal{L}\rangle_{\mathcal{P}_\beta} = \log\mathcal{L}^* \quad (22)$$

C. *Bayesian temperature*

We furthermore propose a temperature that is obtained via Bayesian inference, which returns a distribution rather than a point estimate for β. Since each value of β leads to a different likelihood $\mathcal{L}^\beta$, one can consider the posterior distribution as a function of $\log X$ to be *conditioned* on β. We can therefore write

$$\mathcal{P}(\log X \mid \beta) = \frac{\mathcal{L}^\beta(X)X}{\mathcal{Z}(\beta)}. \quad (23)$$

What we would really like is the distribution of β at the present iteration, so the natural step is to invert this via Bayes' rule;

$$P(\beta \mid \log X^*) = \frac{\mathcal{P}(\log X^* \mid \beta)\,P(\beta)}{P(\log X)}. \quad (24)$$

As with all Bayesian analyses, the distribution of β is fixed up to a prior, which we choose to be uniform in β. The obtained temperatures are consistent with the previous two choices, which may seem oddly coincidental. However, closer inspection reveals that large values of $P(\beta \mid \log X^*)$ are the temperatures with a large value of the posterior at the present contour, normalised by the corresponding evidence. Thus the peak is just the temperature that causes the posterior to be sharply peaked at the current contour i.e. the location of the posterior bulk. Heuristically, it is doing a similar thing as the previous methods.

Comparisons

Unlike annealing, the temperature is not monotonic with progress, which is precisely what allows nested sampling to handle phase transitions.

Free parameter; choose definition based on what you would like to use β to do.

3.4 Dimensionality

We can immediately use the inferred temperature to track how the effective dimensionality of the posterior changes throughout the run, which was previously inaccessible. Handley & Lemos (2019) demonstrated that at the end of a run, a measure of the number of constrained parameters is given by the Bayesian model dimensionality (BMD), defined as the posterior variance of the information content:

$$\frac{\tilde{d}}{2} = \int \mathcal{P}(\theta)\left(\log\frac{\mathcal{P}(\theta)}{\pi(\theta)} - \mathcal{D}_{\mathrm{KL}}\right)^2 \mathrm{d}\theta = \langle \mathcal{I}^2\rangle_{\mathcal{P}} - \langle \mathcal{I}\rangle^2_{\mathcal{P}}. \tag{25}$$

Calculating the quantity using intermediate set of weighted samples (which is concentrated at a single point) leads to vanishing variance, hence also dimensionality. However, we can recover the structure of the posterior together with the true dimensionality by adjusting the temperature. Dimensionality estimates are plotted in fig X for a spherical Gaussian of various dimensions. The Bayesian temperature appears to make the best estimates of the dimension, so that is the one we choose.

Plot of results for spherical Gaussian, for which we know the correct answer.

Parameter compression rate

Plots of samples dimensionality against compression also draw attention to the *speed* at which different parameters are constrained throughout the run. As a concrete example, consider an elongated Gaussian in a unit hypercube prior with $\mu = 0$ and $\Sigma = \mathrm{diag}\,(1, 1, 0.01, 0.01) \times 10^{-2}$.

Plot of dimensionality vs compression factor, showing step changes

It is important to appreciate that at lower compressions the samples truly lie in a lower-dimensional space, because the algorithm has not yet begun to constrain the other parameters. Anticipating the full dimensionality of the space is therefore just as impossible as that associated with a slab-spike geometry, so in this sense such geometries contain a phase transition.

4 ENDPOINT PREDICTION

The time complexity of nested sampling (Petrosyan & Handley 2022) is

$$T \propto n_{\mathrm{live}} \times \langle \mathcal{T}\{\mathcal{L}(\theta)\}\rangle \times \langle \mathcal{T}\{\mathrm{Impl.}\}\rangle \times \mathcal{D}_{\mathrm{KL}}. \tag{26}$$

The second term is the time per likelihood evaluation, which is constant. The third is the cost to replace a dead point with a live point at higher likelihood, which is given by the implementation and usually does not vary in orders of magnitude. The primary unknown during a run, and therefore our primary interest, is the final term: the compression required to get from prior to posterior.

4.1 The termination prior volume

To be exact, one wishes to find the compression factor $\log X_{\mathrm{f}}$ at which the termination criterion is met, which is slightly larger in magnitude than $\mathcal{D}_{\mathrm{KL}}$ (Fig X). The problem is that at an intermediate iteration we only know the posterior up to the maximum log-likelihood live point, which until just before the end is quite far from the posterior bulk.

In order to get an idea of where the true posterior bulk sits, we need to predict what the posterior looks like past the highest live point. We do this 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 predicted posterior to do inference, since more accuracy can always be achieved by simply finishing the run. However, it is more than sufficient for making a prediction for $\log X_{\mathrm{f}}$. Quantitatively, this proceeds as follows: fit a function $f(X, \phi)$ with some parameters ϕ to the known likelihood profile, which allows us to express the prior volume we need to compress to as

$$\Delta\mathcal{Z} = \epsilon\mathcal{Z}_{\mathrm{tot}}, \tag{27}$$

$$\int_0^{X_{\mathrm{f}}} f(X,\phi)\,\mathrm{d}X = \epsilon\left(\int_0^{X_i} f(X,\phi)\,\mathrm{d}X + \mathcal{Z}_{\mathrm{dead}}\right), \tag{28}$$

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 depends 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

$$N_{\mathrm{f}} = -n_{\mathrm{live}} \log X_{\mathrm{f}}. \tag{29}$$

4.2 Regression procedure

A key observation is that the Bayesian model dimensionality is the equivalent dimension of the posterior if it were actually Gaussian. Fitting a Gaussian of this dimension to the likelihood profile therefore makes a reasonable approximation to the true distribution, without explicitly assuming the form of the likelihood function (see plot). The explicit form of the Gaussian that we fit is the same as that given in section X, which we shall repeat here for clarity;

$$f(X;\phi) = \log\mathcal{L}_{\max} - X^{2/d}/2\sigma^2 \tag{30}$$

The extrapolation then proceeds thus:

(i) Find the current dimensionality $\tilde{d}^*$ of the posterior at the Bayesian temperature
(ii) Take the live point profile and do a least squares fit to (30), stipulating that $d = \tilde{d}$ to infer $\log\mathcal{L}_{\max}$ and σ
(iii) Use the likelihood predicted by these parameters to solve (28) for X_{f}

The advantage of fitting a Gaussian is that the procedure can be sped up analytically. Firstly, the least squares regression is trivial because analytic estimators exist; the cost function

$$C^2(\log\mathcal{L}_{\max}, \sigma) = \sum_i \left|\log\mathcal{L}_i - f(X_i; \log\mathcal{L}_{\max}, \sigma)\right|^2 \tag{31}$$

is minimised with respect to $(\log\mathcal{L}_{\max}, \sigma)$ when

$$\sigma^2 = \frac{N\sum_i X_i^{4/d} - \left(\sum_i X_i^{2/d}\right)^2}{2\sum_i \log\mathcal{L}_i \sum_i X_i^{2/d} - 2N\sum_i X_i^{2/d}\log\mathcal{L}_i}, \tag{32}$$

and

$$\log \mathcal{L}_{\max} = \frac{1}{N}\sum_i \log \mathcal{L}_i + \frac{1}{2N\sigma^2}\sum_i X_i^{2/d}. \tag{33}$$

Secondly, the termination prior volume can also be obtained analytically. Rewriting Eq. (28) in terms of the Gaussian parameters gives

$$\epsilon = \frac{\int_0^{X_{\rm f}} \mathcal{L}_{\max} \exp\left(-X^{2/d}/2\sigma^2\right)\, \mathrm{d}X}{\int_0^{X_i} \mathcal{L}_{\max} \exp\left(-X^{2/d}/2\sigma^2\right)\, \mathrm{d}X + \mathcal{Z}_{\rm dead}}. \tag{34}$$

The integrals have the analytic solution

$$\int_0^{X_k} \mathcal{L}_{\max} \exp\left(-X^{2/d}/2\sigma^2\right)\, \mathrm{d}X = \frac{d}{2}\cdot\left(\sqrt{2}\sigma\right)^d \cdot \gamma_k \tag{35}$$

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

$$\log X_{\rm f} = \frac{d}{2}\log 2\sigma^2 + \log \Gamma^{-1}_{d/2}\left(\epsilon\gamma_i + \frac{\epsilon \mathcal{Z}_{\rm dead}}{(2\sigma^2)^{d/2}\,\mathcal{L}_{\max}}\right), \tag{36}$$

and $N_{\rm f}$ is of course just $-n_{\rm live}$ multiplied by this. Intuitively, the above procedure can be thought of as inferring the number of constrained parameters, then extrapolating them up to find the point at which they will be fully constrained.

Plot of fits to Gaussian for different distributions

One might wonder why we do not obtain d via least squares regression together with the other parameters; extensive testing has shown it to be far less stable.

4.3 Alternative approaches

Ultranest (Buchner 2021) tracks progress based on the remaining integral, approximated as $\mathcal{L}_{\max} X_i$. does not correspond directly to runtime, since it will be zero for most of the run before the bulk is crossed.

Extrapolating evidence increments is considerably less stable

5 RESULTS

Now that we have a method for predicting the endpoint, it can be tested on a range of distributions. We begin by considering a series of toy examples to explore the capabilities and limitations of the method, before presenting results for real cosmological chains.

5.1 Toy examples

A. Gaussians

Spherical Gaussian, elongated Gaussian

B. Cauchy

Cauchy

5.2 Cosmological examples

6 CONCLUSION

APPENDIX A: THE MAXIMUM LIVE LOG-LIKELIHOOD

Assume a Gaussian likelihood

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

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

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

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, $\log \mathcal{L}^{\rm live}_{\max}$. 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

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

which may be roughly summarised as

$$y^{\rm live}_{\max} \sim 1 - \frac{\Gamma(1+\frac{2}{d})\Gamma(1+n)}{\Gamma(1+\frac{2}{d}+n)} \tag{A8}$$

$$\pm\left(\frac{\Gamma(1+n)\Gamma(1+\frac{4}{d})}{\Gamma(1+\frac{4}{d}+n)} - \frac{\Gamma(1+\frac{2}{d})^2\Gamma(1+n)^2}{\Gamma(1+\frac{2}{d}+n)^2}\right)^{\frac{1}{2}}, \tag{A9}$$

or in the large d, n limit

$$\lim_{d\to\infty} y^{\rm live}_{\max} \sim \frac{2H_n}{d} \pm \left(\frac{2(\pi^2 - 6\Psi^{(1)}(1+n))}{3d^2}\right)^{\frac{1}{2}}, \tag{A10}$$

$$\lim_{d,n\to\infty} y^{\rm live}_{\max} \sim \frac{2\log n}{d} \pm \sqrt{\frac{2}{3}}\frac{\pi}{d}, \tag{A11}$$

where $\psi^{(1)}$ is the trigamma function and H_n is the nth harmonic number.

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.

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
Buchner J., 2021, UltraNest – a robust, general purpose Bayesian inference engine (arXiv:2101.09604)
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
Habeck M., 2015, AIP Conference Proceedings, 1641, 121
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
Higson E., Handley W., Hobson M., Lasenby A., 2019, Statistics and Computing, 29, 891–913
Petrosyan A., Handley W., 2022, Physical Sciences Forum, 5
Skilling J., 2006, Bayesian Analysis, 1