

aeons: approximating the end of nested sampling

Zixiao Hu^{*,1,2} Artem Baryshnikov^{1,2} Will Handley^{†1,2}

¹*Astrophysics Group, Cavendish Laboratory, J. J. Thomson Avenue, Cambridge, CB3 0HE, UK*

²*Kavli Institute for Cosmology, Madingley Road, Cambridge, CB3 0HA, UK*

27 November 2023

ABSTRACT

This paper presents analytic results on the anatomy of nested sampling, from which a technique is developed to estimate the run-time of the algorithm that works for any nested sampling implementation. We test these methods on both toy models and true cosmological nested sampling runs. The method gives an order-of-magnitude prediction of the end point at all times, forecasting the true endpoint within standard error generally around the halfway point.

Key words: methods: data analysis – methods: statistical

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 (Trotta et al. 2008), materials science (Pártay et al. 2021) and machine learning (Higson et al. 2018a). The core algorithm is unique in its estimation of volumes by *counting*, coupling together nested monte carlo integrals which makes high-dimensional integration feasible and robust. It also avoids problems that challenge traditional Bayesian samplers, such as posterior multi-modality and phase transitions.

The order of magnitude run-time 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 (e.g. Feroz et al. 2009; Handley et al. 2015; Brewer & Foreman-Mackey 2016; Speagle 2020; Buchner 2021; Williams et al. 2021; McEwen et al. 2023) either do not give an indication of remaining run-time, or only provide crude measures of progress that do not directly correspond to the true endpoint.

This paper sets out a principled manner of endpoint estimation for nested sampling at each intermediate stage (as shown in Fig. 1), the key idea being to use the existing samples to predict the likelihood in the region we have yet to sample from. We begin with an overview of nested sampling in Section 2, followed by an examination of the anatomy of a nested sampling run to establish key concepts for endpoint prediction in Section 3. Section 4 then outlines the methodology we use, including discussion and comparisons to previous attempts. Finally, Section 5 presents the results and discussions for toy and cosmological chains, before we conclude.

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 BACKGROUND

We begin with a brief description of the nested sampling algorithm to establish the necessary notation. For a more comprehensive treatment, we recommend the original (Skilling 2006) paper, the Sivia & Skilling (2006) textbook, as well as the excellent technical review by Buchner (2023) and Nature review by Ashton et al. (2022).

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

$$\mathcal{Z} = \int \mathcal{L}(\theta) \pi(\theta) d\theta \quad (1)$$

while producing samples of the posterior distribution

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

The algorithm operates by maintaining a set of n_i *live points* sampled from the prior, which can vary in number throughout the run (Higson et al. 2019). At each iteration i , the point with the lowest likelihood is removed and added to a list of *dead points* (illustrated in Fig. 2). New points are then (optionally) drawn from the prior, subject to the constraint that they must have a higher likelihood than the latest dead point \mathcal{L}_i . Repeating the procedure leads to the live points compressing around peaks in the likelihood. If there are more births than deaths, the number of live points n_i increases, whilst choosing to not generate new points reduces the number of live points. The exact creation schedule depends on the dynamic nested sampling strategy.

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

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

* E-mail: zixiao.hu.09@gmail.com

† E-mail: wh260@cam.ac.uk

where $X(\mathcal{L})$ is the fraction of the prior with a likelihood greater than \mathcal{L} . Whilst the likelihood contour \mathcal{L}_i at each iteration is known, the prior volumes X_i must 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

$$X_i = \prod_{k=1}^i t_k. \quad (4)$$

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

$$P(t_i) = n_i t_i^{n_i-1}, \quad (5)$$

$$\langle \log t_i \rangle = -\frac{1}{n_i}, \quad \text{Var}(\log t_i) = \frac{1}{n_i^2}. \quad (6)$$

The algorithm terminates when an user-specified condition is met; a popular choice is when the evidence in the live points falls below some fraction ϵ of the accumulated evidence e.g. 10^{-3} , which is proven to be a valid convergence criterion (Evans 2006; Chopin & Robert 2010). Much of the existing literature treats this remaining evidence separately, for instance by estimating it as the termination X multiplied by the average likelihood amongst the remaining live points. It is, however, quantitatively equivalent but qualitatively neater to consider termination as killing the remaining live points off one-by-one, incrementing the evidence exactly as during the run with decreasing n_i (Speagle 2020).

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 \mathbf{t} . As Skilling and others (Chopin & Robert 2010; Keeton 2011) have shown, for any given problem the uncertainty in $\log \mathcal{Z}$ is proportional to $1/\sqrt{n_{\text{live}}}$, so n_{live} sets the resolution of the algorithm.

3 THE ANATOMY OF A NESTED SAMPLING RUN

The following sections act as an inventory of the information available to us at an intermediate iteration i , which we shall use to make endpoint predictions in Section 4. We present an anatomy of the progression of a nested sampling run in terms of the prior volume compression (Section 3.1), the log-likelihood increase (Section 3.2), the inferred temperature (Section 3.3), and the dimensionality of the samples (Section 3.4).

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*:

$$\mathcal{D}_{\text{KL}} = \int \mathcal{P}(\theta) \log \frac{\mathcal{P}(\theta)}{\pi(\theta)} d\theta. \quad (7)$$

The bulk of the posterior lies within a prior volume $X = e^{-\mathcal{D}_{\text{KL}}}$, which is the target compression. From Eq. (6) one gets there by iteratively taking steps of size

$\Delta \log X_i = -1/n_i$, so that when we add up the contribution of each step in (6) we get

$$\langle \log X_i \rangle = -\sum_{k=1}^i \frac{1}{n_k}, \quad \text{Var}(\log X_i) = \sum_{k=1}^i \frac{1}{n_k^2}. \quad (8)$$

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

The same is not true for the live points, which are uniformly distributed in prior volume, and by the comments in the penultimate paragraph of Section 2 have $n_{i+k} = n_i - k$. As a result, the maximum live point is found at

$$\langle \log X_{\text{min}}^{\text{live}} \rangle = \langle \log X_i \rangle - \sum_{k=1}^{n_i} \frac{1}{n_{i+k}} \approx -\frac{i}{n_i} - \log n_i - \gamma, \quad (9)$$

with variance

$$\text{Var}(\log X_{\text{min}}^{\text{live}}) = \text{Var}(\log X_i) + \sum_{k=1}^{n_i} \frac{1}{n_{i+k}^2} \approx \frac{i}{n_i^2} + \frac{\pi^2}{6}, \quad (10)$$

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

The live points therefore only get us a factor of $\log n_i$ closer in volume to the posterior bulk. In other words, it is not until we are around $\log n_i$ away from $\log X = -\mathcal{D}_{\text{KL}}$ that the samples begin populating the posterior typical set. One can see from (7) 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.

The result is consistent with that in Pártay et al. (2010), which states that a spike at a volume smaller than X_i/n_i will go undetected. 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. These results are summarised in Fig. 3

3.2 Log-likelihood

We now consider the distribution of the live and dead points in log-likelihood. We will examine the representative case of the d -dimensional multivariate Gaussian:

$$\log \mathcal{L} = \log \mathcal{L}_{\text{max}} - X^{2/d}/2\sigma^2 \quad (11)$$

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

$$\langle \log \mathcal{L} \rangle_{\mathcal{P}} = \log \mathcal{L}_{\text{max}} - \frac{d}{2}, \quad \text{Var}(\log \mathcal{L})_{\mathcal{P}} = \frac{d}{2}. \quad (12)$$

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

$$y = \frac{\log \mathcal{L} - \log \mathcal{L}_i}{\log \mathcal{L}_{\text{max}} - \log \mathcal{L}_i}; \quad (13)$$

$y = 0$ corresponds to the current point $\log \mathcal{L}_i$ and $y = 1$ to the maximum $\log \mathcal{L}_{\text{max}}$. At each iteration, y increases by roughly

$$\lim_{n_i \rightarrow \infty} \Delta y \approx \frac{dy}{d \log X} \Delta \log X = \frac{2}{dn_i} \quad (14)$$

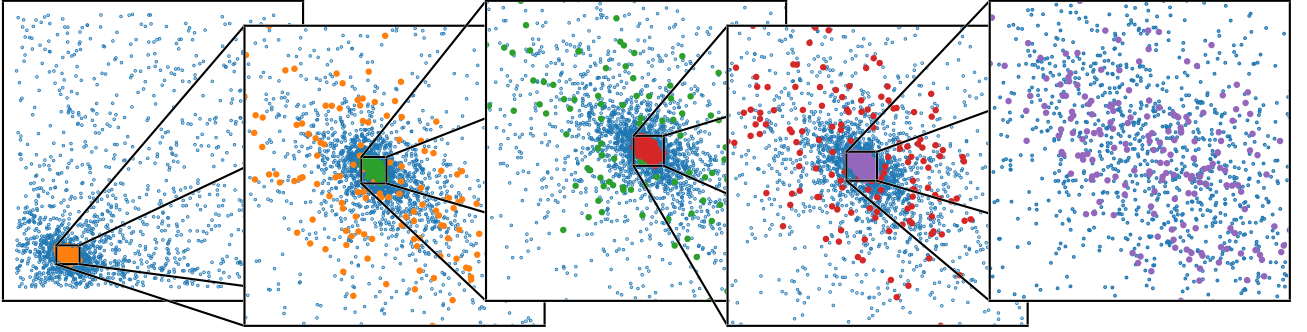


Figure 2. The dead points of a nested sampling run, recursively zoomed in. Their density is constant in $\log X$, which is a geometrically constant measure $\propto dX/X$, and hence scale invariant until the final set of live points is reached. The live points at a given iteration i (larger dots) are uniform across the prior, plotted for comparison.

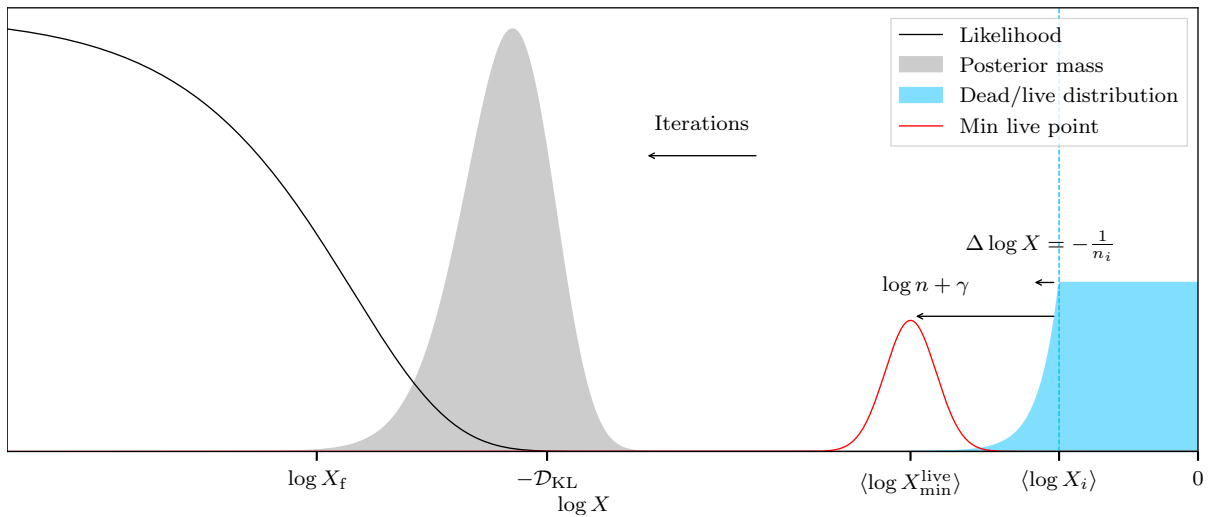


Figure 3. The distribution of the posterior mass in terms of $\log X$, the live points over the constrained prior and the smallest live point prior volume $\log X_{\min}^{\text{live}}$ at an intermediate iteration i . For large values of \mathcal{D}_{KL} i.e. informative posteriors and/or large dimensionalities, 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 onward will be in terms of $-\log X$, where the run will more naturally proceed from left to right.

in the large n_i limit. We can again sum the harmonic series to see that the maximum live point is expected to be at

$$y_{(n_i)}^{\text{live}} = \frac{2 \log n_i}{d}, \quad (15)$$

where we have used the notation of order statistics to denote $x_{(n)}$ as the maximum of n points. This is 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

$$y_{(n_i)}^{\text{live}} - y_{(n_i-1)}^{\text{live}} \approx \frac{2}{d}. \quad (16)$$

Before reaching the posterior bulk, $\log \mathcal{L}_{\max} - \log \mathcal{L}_i > d/2$, so we must have

$$\log \mathcal{L}_{(n_i)}^{\text{live}} - \log \mathcal{L}_{(n_i-1)}^{\text{live}} > 1. \quad (17)$$

In other words, the highest likelihood point is always at least

an order of magnitude greater than the second highest. 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 on a more quantitative footing.

A more rigorous derivation (Appendix A) shows that the

maximum live point has mean and variance

$$\lim_{d, n_i \rightarrow \infty} y_{\max}^{\text{live}} \sim \frac{2 \log n_i}{d} \pm \sqrt{\frac{2}{3}} \frac{\pi}{d}. \quad (18)$$

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

$$\epsilon = \frac{\int_0^{X_f} \mathcal{L} dX}{\int_0^\infty \mathcal{L} dX}. \quad (19)$$

Note that we have assumed that prior effects are negligible (so $1 \approx \infty$), and that $\epsilon \ll 1$ so that the denominator is approximately the accumulated evidence. Computing this for (11), we find the answer in terms of lower incomplete gamma functions

$$\epsilon = 1 - \frac{\Gamma_{d/2} \left(X_f^{2/d} / 2\sigma^2 \right)}{\Gamma(d/2)}. \quad (20)$$

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

$$\lim_{X_f \ll (\sqrt{2}\sigma)^d} \epsilon \approx \frac{X_f}{(\sqrt{2}\sigma)^d \Gamma(1 + \frac{d}{2})} = \frac{(\log \mathcal{L}_{\max} - \log \mathcal{L}_f)^{\frac{d}{2}}}{\Gamma(1 + \frac{d}{2})}. \quad (21)$$

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

$$\lim_{d \rightarrow \infty} \log \mathcal{L}_f \approx \log \mathcal{L}_{\max} - \frac{d}{2e}. \quad (22)$$

In the event that we keep ϵ in, we replace $\frac{d}{2e} \rightarrow \frac{d}{2e} \epsilon^{2/d}$, 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}_i in (13) to be \mathcal{L}_f , and combining this with (18) we find

$$\log \mathcal{L}_{\max}^{\text{live}} \approx \log \mathcal{L}_{\max} - \frac{d}{2e} + \frac{\log n}{e} \pm \frac{\pi}{\sqrt{6}e}, \quad (23)$$

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 $\log n/e$ closer, with a chance of getting $\sim \pi/\sqrt{6}e = 0.472$ closer still by statistical fluctuation.

A plot of the log-likelihood distribution at the end of a run is shown in Fig. 4

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, so we can transform the likelihood as $\mathcal{L} \rightarrow \mathcal{L}^\beta$ without loss of information by trivially re-weighting the samples. Increasing β worsens the situation, while $\beta \rightarrow 0$ simply gives back the prior. There is, on the other hand, a significant intermediate range which makes the samples look like a posterior centred at the present

contour, which will allow us to recover the structure of the samples. A schematic of the procedure is shown in Fig. 5.

At this point it is relevant to note the correspondence between Bayesian inference and statistical mechanics, from which the above transform is derived. If one equates the parameters to microstates i , the negative log-likelihood to the microstate energy E_i , and the prior to the density of states g_i , then the posterior as given by the generalised Bayes' rule is the canonical ensemble

$$p(E_i) = \frac{g_i e^{-\beta E_i}}{Z(\beta)} \leftrightarrow \mathcal{P}_\beta(\theta) = \frac{\mathcal{L}^\beta(\theta) \pi(\theta)}{\mathcal{Z}(\beta)} \quad (24)$$

at the inverse temperature $\beta = 1/T$. As noted by Habeck (2015), thermal algorithms such as thermodynamic integration (Gelman & Meng 1998) get the evidence by evolving through a series of canonical ensembles via some temperature schedule, but nested sampling instead maintains *microcanonical* ensembles, which are the iso-likelihood contours. Instead of using temperature (Kirkpatrick et al. 1983; Swendsen & Wang 1986) or energy (Wang & Landau 2001) as a control parameter, nested sampling chooses a series of ensembles with constant relative volume entropy $\Delta \log X$, which allows the algorithm to handle phase transitions (Baldock et al. 2016).

Because the temperature of a microcanonical ensemble is a derived property rather than a parameter, there is some freedom in its definition. Returning to our original motivation, we make the connection that the temperature is the re-weighting $\mathcal{L} \rightarrow \mathcal{L}^\beta$ which centres the ensemble around the current energy. We now present several temperatures that achieve this aim, each of which one can plausibly consider to be the current temperature of a nested sampling run.

A. Microcanonical temperature

The obvious candidate is the microcanonical temperature $\partial S / \partial E$, where the volume entropy is $\log X$ and the energy is as usual $-\log \mathcal{L}$. This gives the density of states; as discussed in Skilling's original paper,

$$\beta_M = - \left. \frac{d \log X}{d \log \mathcal{L}} \right|_{\log \mathcal{L}_i}. \quad (25)$$

is the β at which $\mathcal{L}^\beta X$ peaks at $\log X_i$, if we assume differentiability, which is exactly the intuition we were aiming for to put the ensemble bulk at the current contour.

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. Indeed, material science applications (Baldock et al. 2017) use this estimator to monitor the 'cooling' progress of nested sampling, with a window size of 1000 iterations.

B. Canonical temperature

Another temperature considered by Habeck (2015) is that at which the current energy (i.e. $-\log \mathcal{L}_i$) is the average energy of the entire ensemble. One can obtain it by inverting

$$\langle \log \mathcal{L} \rangle_{\mathcal{P}_\beta} = \log \mathcal{L}_i \quad (26)$$

to get the 'canonical' temperature β_C . While β_M is derived from (the gradient of) a single contour, this temperature uses the entire ensemble. It has the desirable property that it rises

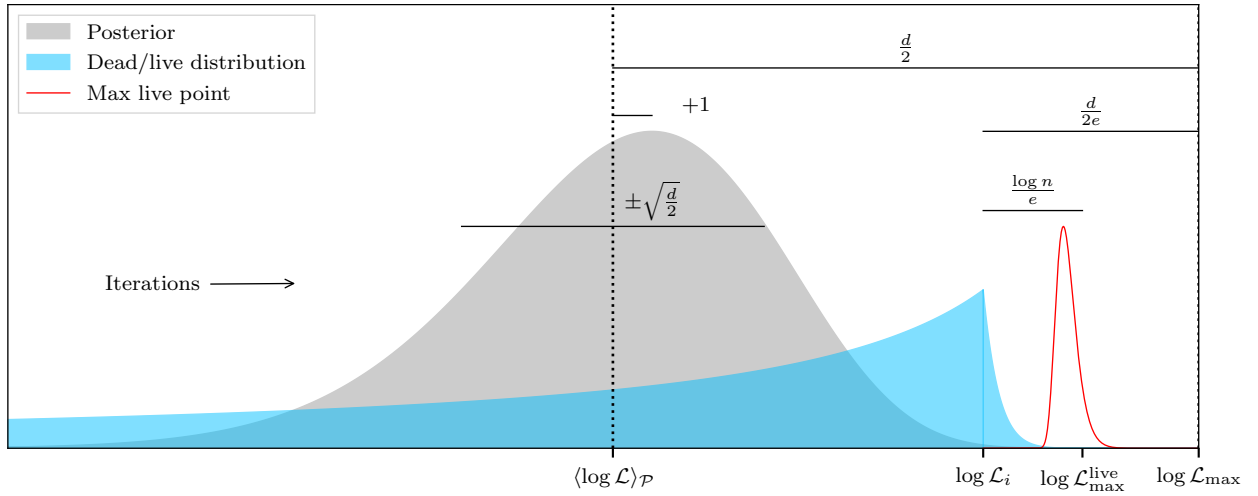


Figure 4. Distribution of samples as a function of $\log \mathcal{L}$, showing the posterior $\mathcal{P}(\log \mathcal{L})$, the distribution of the live points $\pi(\log \mathcal{L} | \mathcal{L} > \mathcal{L}_i)$, and the distribution of the maximum likelihood live point $P(\log \mathcal{L}_{\max}^{\text{live}})$. The distances are shown between these locations at the end of the run, the key takeaway being that in high dimensions the highest log-likelihood point of a nested sampling run is nowhere near the maximum in high dimensions.

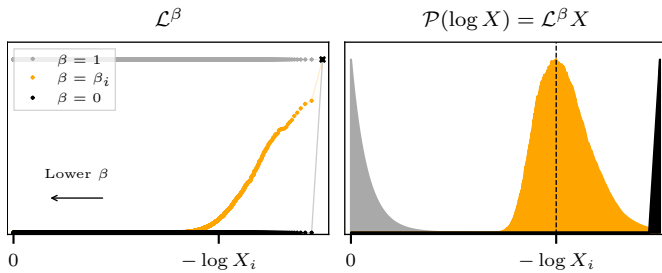


Figure 5. The likelihood and posterior as a function of $\log X$ in the middle of a nested sampling run. Almost all of the posterior mass is concentrated at a single point with the highest compression, because it is orders of magnitude higher in likelihood. Reducing β re-weights to shift the posterior mass; taking $\beta \rightarrow 0$ goes too far and gives back the prior, but there is some intermediate $\beta = \beta_i$ which is significant because it centres the posterior at the current contour.

monotonically with compression, in analogy to a monotonic annealing schedule.

C. Bayesian temperature

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

$$\mathcal{P}(\log X | \beta) = \frac{\mathcal{L}^\beta(X)X}{\mathcal{Z}(\beta)}. \quad (27)$$

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;

$$P(\beta | \log X_i) = \frac{\mathcal{P}(\log X_i | \beta) P(\beta)}{P(\log X)}. \quad (28)$$

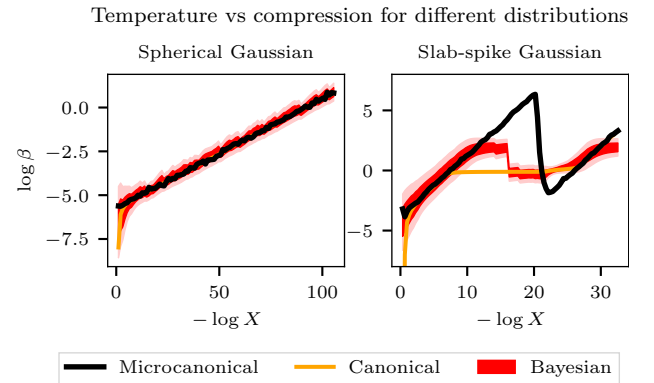


Figure 6. Inferred temperatures using the microcanonical, canonical and Bayesian definitions. The shaded regions show the $1 - 2\sigma$ uncertainties. All are consistent for a single phase, but differ during a phase transition.

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. In fact, closer inspection reveals that large values of $P(\beta | \log X_i)$ are the temperatures with a large value of the posterior at the present contour, normalised by the corresponding evidence. Thus the Bayesian temperature uses the same idea as the microcanonical one, except it accounts for the spread in the result.

Comparisons

Fig. 6 shows the three temperatures as a function of compression for two cases, one containing a phase transition and one without. They are consistent in both cases when there is a single dominant phase, but differ during a phase transition. The canonical temperature is the only one that rises monotonically with compression.

One should keep in mind that despite the above theoretical

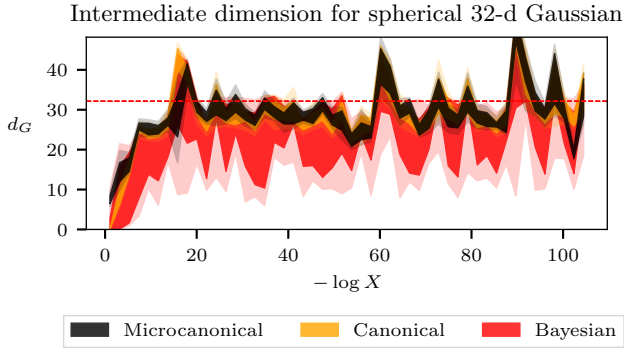


Figure 7. Dimensionality estimates using the different temperatures for a spherical 32-d Gaussian. Again, all are consistent, but the Bayesian definition has an uncertainty which includes the true value far more consistently than for the other definitions.

reasoning, our introduction of the likelihood transformation was ultimately motivated by our wish to utilise the extra degree of freedom it provides. As we will see below, we recommend choosing the exact definition depending on what is useful for the problem at hand.

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:

$$\frac{d_G}{2} = \int \mathcal{P}(\theta) \left(\log \frac{\mathcal{P}(\theta)}{\pi(\theta)} - \mathcal{D}_{\text{KL}} \right)^2 d\theta = \langle \mathcal{I}^2 \rangle_{\mathcal{P}} - \langle \mathcal{I} \rangle_{\mathcal{P}}^2. \quad (29)$$

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. 7 for a spherical 32-d Gaussian, for which the true dimensionality is known.

The different choices of temperature are again consistent, but for the rest of this paper we choose the Bayesian β , because it provides a better reflection of the uncertainty in the estimate; the others, while fluctuating around the true value, are often many standard errors away from the true value at each single point.

Anisotropic compression

Plots of samples dimensionality against compression also draw attention to the *directions* in which the samples are constrained throughout the run. As a concrete example, consider an elongated Gaussian in a unit hypercube prior with $\mu = \mathbf{0}$ and $\Sigma = \text{diag}(10^{-3}, 10^{-3}, 10^{-3}, 10^{-6}, 10^{-6}, 10^{-6})$, for which the dimensionality estimates are plotted in Fig. 8. Alongside is a view of the distribution of live points across the prior for two directions with different scales, which shows the level

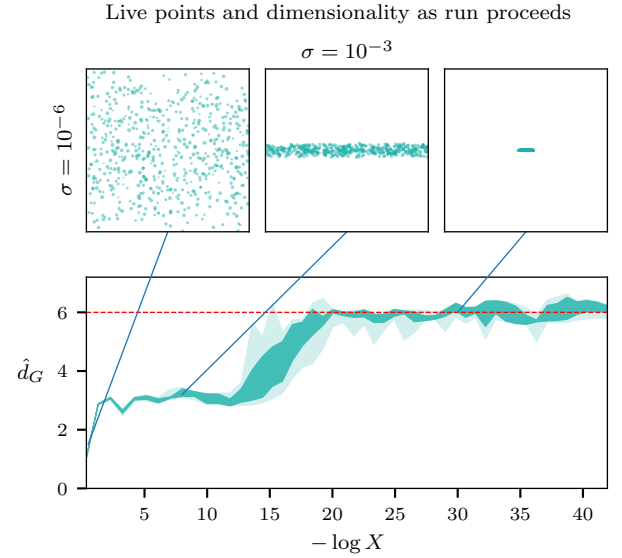


Figure 8. Dimensionality estimates for a Gaussian that is elongated in half of its dimensions, with $1 - 2\sigma$ uncertainties shaded. The locations of the live points in the prior are shown at three stages, indicated by the connecting lines. As can be seen from the live point distribution, the prior does not compress in the higher variance direction until much later in the run, and early on it appears as if those directions are completely unconstrained.

to which those parameters have been constrained at different times.

A feature of nested sampling made apparent here 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, and initially it is much easier to find a better point along the direction of a parameter that is poorly constrained. Lower variance parameters are constrained much later, and before that happens it appears as though those parameters have a uniform distribution.

It is important to appreciate that at lower compression the samples truly lie in a lower-dimensional space, rather than some artefact of the way we view them. 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 ‘compressive phase transition’.

4 ENDPOINT PREDICTION

As described in Petrosyan & Handley (2022) and further explored in the talk and upcoming paper Handley (2023a,b), the time complexity of nested sampling is

$$T \propto \langle \frac{1}{n_i} \rangle^{-1} \times \langle \mathcal{T}\{\mathcal{L}(\theta)\} \rangle \times \langle f_{\text{sampler}} \rangle \times \mathcal{D}_{\text{KL}}. \quad (30)$$

The first term is the harmonic mean of the number of live points $\sim \mathcal{O}(n)$, The second term is the average time per likelihood evaluation. The third is the average number of evaluations required 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 final term is the Kullback-Liebler divergence Eq. (7), the compression factor required to get from the prior to the posterior. This

term is generally outside of user control, in most cases *a priori* unknown, and of principle interest in this section.

4.1 The termination prior volume

Making the above discussion more precise, we wish to find the compression factor $\log X_f$ at which the termination criterion is met, which is larger in magnitude than \mathcal{D}_{KL} (Fig. 3). The difficulty 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 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 perform inference, since more accuracy can always be achieved by simply finishing the run. However, we will demonstrate it is sufficient for making a run-time prediction for $\log X_f$.

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

$$\Delta \mathcal{Z} = \epsilon \mathcal{Z}_{\text{tot}}, \quad (31)$$

or equivalently

$$\int_0^{X_f} f(X, \phi) dX = \epsilon \left(\int_0^{X_i} f(X, \phi) dX + \mathcal{Z}_{\text{dead}} \right), \quad (32)$$

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

Once X_f is known, the corresponding iteration count depends on the live point schedule. For example in the constant n_i case at each iteration $\log X$ decreases by $1/n$, so the total number of iterations N_f would be

$$N_f = -n \log X_f. \quad (33)$$

4.2 How to extrapolate?

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. The parameterisation of the Gaussian that we fit is the same as that given in Section 3.2, which we shall repeat here for clarity;

$$f(X; \phi) = \log \mathcal{L}_{\text{max}} - X^{2/d}/2\sigma^2 \quad (34)$$

The extrapolation then proceeds thus:

- (i) Find the current dimensionality $d_G^{(i)}$ of the posterior at the Bayesian temperature
- (ii) Take the live point profile and perform a least squares fit to (34), stipulating that $d = d_G^{(i)}$ to infer $\log \mathcal{L}_{\text{max}}$ and σ
- (iii) Use the likelihood predicted by these parameters to solve (32) 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

$$C^2(\log \mathcal{L}_{\text{max}}, \sigma) = \sum_i |\log \mathcal{L}_i - f(X_i; \log \mathcal{L}_{\text{max}}, \sigma)|^2 \quad (35)$$

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

$$\sigma^2 = \frac{N \sum_i X_i^{4/d} - \left(\sum_i X_i^{2/d} \right)^2}{2 \sum_i \log \mathcal{L}_i \sum_i X_i^{2/d} - 2N \sum_i X_i^{2/d} \log \mathcal{L}_i}, \quad (36)$$

and

$$\log \mathcal{L}_{\text{max}} = \frac{1}{N} \sum_i \log \mathcal{L}_i + \frac{1}{2N\sigma^2} \sum_i X_i^{2/d}. \quad (37)$$

Secondly, the termination prior volume can also be obtained analytically. Rewriting Eq. (32) in terms of the Gaussian parameters gives

$$\epsilon = \frac{\int_0^{X_f} \mathcal{L}_{\text{max}} \exp(-X^{2/d}/2\sigma^2) dX}{\int_0^{X_i} \mathcal{L}_{\text{max}} \exp(-X^{2/d}/2\sigma^2) dX + \mathcal{Z}_{\text{dead}}}. \quad (38)$$

The integrals have the analytic solution

$$\int_0^{X_k} \mathcal{L}_{\text{max}} \exp(-X^{2/d}/2\sigma^2) dX = \frac{d}{2} \cdot (\sqrt{2}\sigma)^d \cdot \gamma_k \quad (39)$$

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

$$\log X_f = \frac{d}{2} \log 2\sigma^2 + \log \Gamma_{d/2}^{-1} \left(\epsilon \gamma_i + \frac{\epsilon \mathcal{Z}_{\text{dead}}}{(2\sigma^2)^{d/2} \mathcal{L}_{\text{max}}} \right), \quad (40)$$

and N_f is of course just $-n$ 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.

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

4.3 Alternative approaches

More comprehensive Bayesian approaches, perhaps including *a priori* information about the likelihood or greater flexibility in the fitting function, could likely perform better than what we have just presented. However, such methods would not befit *run-time* prediction which has a much more limited computational budget, hence the more pragmatic approach we have adopted. Here, we discuss as a benchmark alternative approaches to endpoint estimation that have a comparable computational complexity.

A. Integral progress

An alternative approach used in ULTRANEST (Buchner 2021) derives a progress bar based on the fraction of the accumulated integral compared to the remaining integral, approximated as

$$\mathcal{Z}_{\text{rem}} \approx \mathcal{L}_{\text{max}}^{\text{live}} X_i. \quad (41)$$

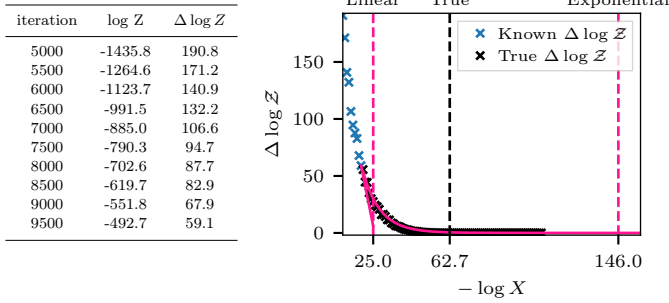


Figure 9. Extrapolating the increments of evidence. The left column shows the output of a nested sampling run, and the right column shows the extrapolation.

This has several shortcomings. First, run-time is proportional to compression rather than accumulation of the integral, since it takes just as long to traverse the width of the bulk as it does any other width. Second, because of the point-like nature of the posterior mid-run, the remaining integral approximated as such holds nearly all of the evidence, so the relative fraction of the accumulated and remaining evidence is almost zero for most of the run. Finally, approximation (41) is always an underestimate, because as previously found the maximum live point is generally nowhere near the true maximum. This approach can however be useful in the low dimensions appropriate for ULTRANEST when the live points are always near the maximum, but in general is less reliable.

B. Extrapolating evidence increments

Seasoned watchers of nested sampling runs 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 and compare it 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. 9. 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 exactly exponential, since the exponential fit leads to a large over-prediction. The predicted endpoint over the course of a run for $d = 16$, $\sigma = 0.01$, as shown in Fig. 10, 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 RESULTS

We now test the approach established in the previous section on a range of distributions. We begin by considering a series of toy examples to explore the capabilities and limitations of

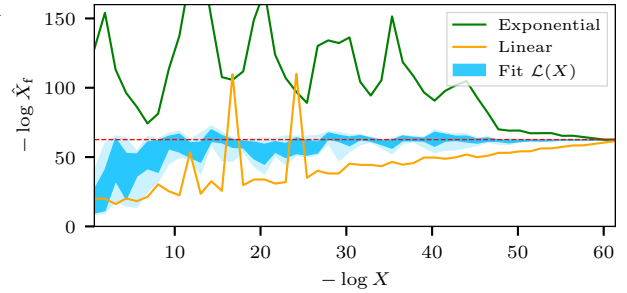


Figure 10. Endpoint predictions for a spherical Gaussian, with $1-2\sigma$ uncertainties shaded. Extrapolating the evidence increments in a linear/exponential manner under/over-predicts the endpoint, both of which perform considerably worse than the method of extrapolating the likelihood.

the method, before presenting results for real cosmological chains.

5.1 Toy examples

Throughout the toy examples we make use of the perfect nested sampling (Keeton 2011; Higson et al. 2018b) framework implemented in ANESTHETIC Handley (2019).

A. Gaussians

Predictions for spherical Gaussians of various dimensions are shown in Fig. 11 as a benchmark for when fitting a Gaussian distribution is exact. In all endpoint prediction plots, the shading indicates the $1-2\sigma$ uncertainties. All were run with $n_i = 500$ except for one $n_i = 2000$ for comparison, with each Gaussian having a width of $\sigma = 0.01$. The correct endpoint is recovered to within standard error at all points except the very beginning, when the parameters have hardly been constrained.

We note that as with nested sampling in general, increasing n_i improves the resolution and reliability of the inferences, which can be seen from the middle two plots. We also observe the effect of elongating the Gaussian, using the same example as Section 3.4. Fig. 12 shows a step-like trend similar to the inferred dimensionalities, reflecting the fact that the full dimensionality is undetectable at lower compression factors. The endpoint for a likelihood whose remaining three directions are completely unconstrained coincides with our predictions at early iterations, showing that the two cases are indistinguishable.

B. Cauchy

One case that might be expected to cause problems is the pathological Cauchy distribution, which is far from a Gaussian. Fig. 13 shows the predictions for a likelihood of the form

$$\log \mathcal{L} = \log \mathcal{L}_{\max} - \frac{1+d}{2} \log \left(1 + \frac{X^2}{\gamma^2} \right), \quad (42)$$

choosing $d = 10$ and allowing γ to vary. The correct estimate is obtained to within standard error by about halfway, but before that is inaccurate. The key limitation is that the estimate is wrong early on, not because the compression is anisotropic, or that there is a phase transition; but rather as

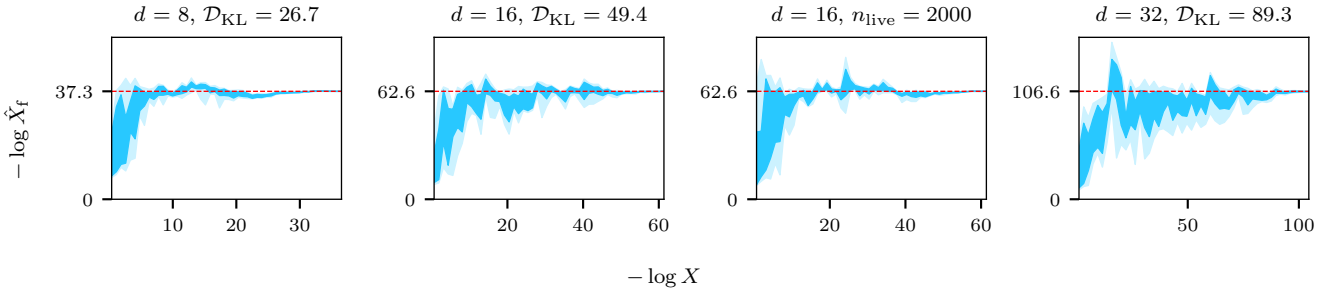


Figure 11. Endpoint predictions for a spherical Gaussian run with $n_i = 500$ (except for the third plot from left). The correct endpoint is obtained for all but the earliest iterations, and the uncertainty is controlled by the number of live points, which can be seen from the two $d = 16$ plots.

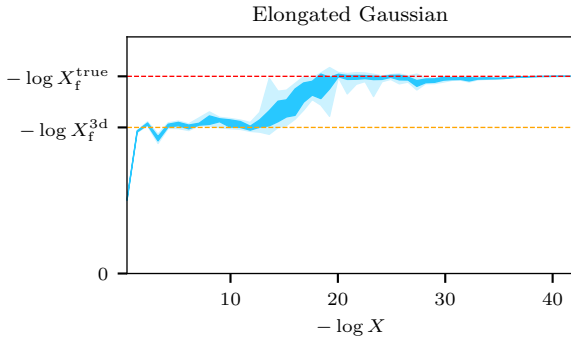


Figure 12. Endpoint prediction for an elongated Gaussian. At early stages, the full dimensionality is undetectable, and the endpoint is predicted to be the same as for a likelihood with three unconstrained directions. Only once the prior has been compressed enough to constrain the other three directions does the prediction converge to the true value.

a limitation of the reducing the likelihood to a Gaussian via the BMD, which is itself less stable for a Cauchy.

Nevertheless, the right order of magnitude is obtained at all times, so this remains sufficient for most use-cases. The Cauchy is also a pathological case, and the same problem does not in practice appear for more realistic cases, as we shall see next.

5.2 Cosmological examples

Finally, we evaluate the method on real cosmological chains. Fig. 14 presents the endpoints (calculated after the fact) for nested sampling runs for curvature quantification on several common cosmological data sets (details in Handley (2021)).

The SH0ES, BAO and LENSING chains are ‘easy’ low \mathcal{D}_{KL} inferences, so it is expected that the correct endpoint is inferred practically from the start. The PLANCK endpoints, on the other hand, are not correct until at least midway through. However, this is expected from the covariance of the PLANCK likelihood, which consists of principal components of many scales and therefore elongated in many dimensions. It is therefore of the same class as the elongated Gaussian presented in Section 3.4; the samples exist in a lower dimensional subspace mid-run, which slowly increases to the full dimensionality only at the end of the run.

6 CONCLUSION

We have derived new analytic results to make an anatomy of nested sampling, understanding the progression of a run via the compression of prior volume, the increase in log-likelihood, the inferred temperature schedule, and the convergence of the sample dimensionality. From these analyses, we developed a method for predicting the endpoint of a nested sampling run, by using the inferred Bayesian model dimensionality mid-run to extrapolate the known likelihood profile.

The method in general converges on a correct prediction of endpoint by about halfway, and gets the correct order of magnitude throughout. Consistent predictions are obtained for both toy and cosmological examples. The accuracy is typically limited by the information available mid-run, either because of a phase transition or because the anisotropy of the nested sampling compression. Pathological distributions, such as a Cauchy, lead to less stable inferences of the dimensionality and expose the limitations of a Gaussian approximation, though the order of magnitude is still correct.

Further work can be done to experiment with more flexible basis functions for regression of the likelihood profile, so that it is less dependent on the Gaussian approximation.

CODE AVAILABILITY

A package implementing the endpoint prediction mechanism is in development for easy plug-in to existing implementations of nested sampling. The latest updates can be found at github.com/zixiao-h/aeons.

ACKNOWLEDGEMENTS

WJH is supported by the Royal Society as a Royal Society University Research Fellow at the University of Cambridge. AB was supported by a Cambridge Mathematical Placement and the Royal Society summer studentship. ZH was supported by a Royal Society summer studentship. We thank Mike Hobson for helpful discussion during ZH Part III viva.

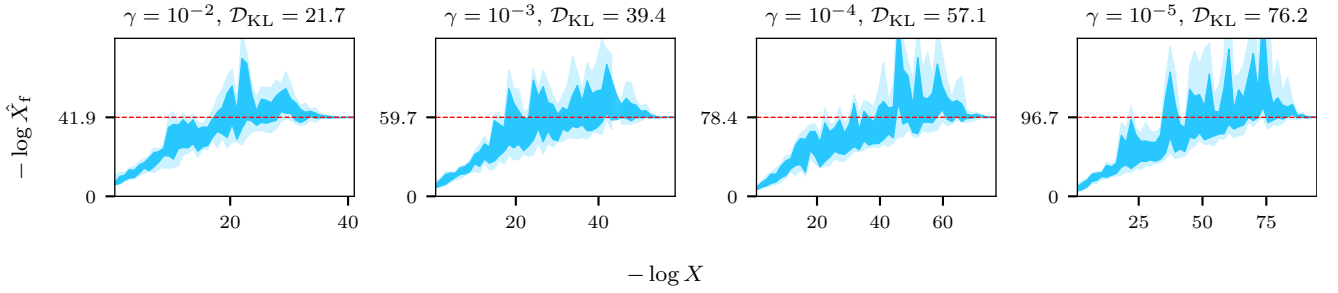


Figure 13. Predictions for the Cauchy distribution for various widths γ and $n_i = 500$. The endpoint is underestimated for the first half of the run, but this is a limitation of the Gaussian approximation rather than a lack of information mid-run.

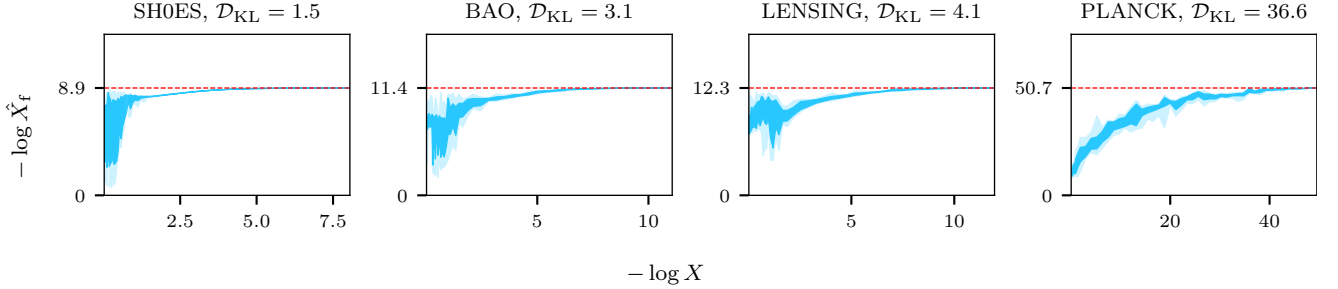


Figure 14. Endpoint predictions for cosmological likelihoods. The first three low \mathcal{D}_{KL} inferences get the correct endpoint from the start, while the PLANCK chain takes longer to converge because the likelihood is highly covariant and thus subject to anisotropic compression, which makes the samples lie in a lower dimensional subspace mid-run.

APPENDIX A: THE MAXIMUM LIVE LOG-LIKELIHOOD

Assume a Gaussian likelihood

$$\log \mathcal{L} = \log \mathcal{L}_{\max} - X^{2/d}/2\sigma^2. \quad (\text{A1})$$

Providing that $\sigma^d \ll 1$, the posterior can be approximated as

$$P(X) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-X^{2/d}/2\sigma^2}. \quad (\text{A2})$$

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

$$P(\log \mathcal{L}) = \frac{1}{\Gamma(\frac{d}{2})} e^{\log \mathcal{L} - \log \mathcal{L}_{\max}} (\log \mathcal{L}_{\max} - \log \mathcal{L})^{\frac{d}{2}-1} \quad (\text{A3})$$

i.e. $2(\log \mathcal{L}_{\max} - \log \mathcal{L}) \sim \chi_d^2$. The posterior average and variance of $\log \mathcal{L}$ are given by

$$\langle \log \mathcal{L} \rangle_{\mathcal{P}} = \log \mathcal{L}_{\max} - \frac{d}{2}, \quad \text{Var}(\log \mathcal{L})_{\mathcal{P}} = \frac{d}{2}. \quad (\text{A4})$$

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

$$P(\log \mathcal{L}) = \frac{d}{2} \frac{(\log \mathcal{L}_{\max} - \log \mathcal{L})^{\frac{d}{2}-1}}{(\log \mathcal{L}_{\max} - \log \mathcal{L}_i)^{\frac{d}{2}}} \quad [\log \mathcal{L}_i < \log \mathcal{L} < \log \mathcal{L}_{\max}], \quad (\text{A5})$$

It is helpful at this stage to define a parameter

$$y = \frac{\log \mathcal{L} - \log \mathcal{L}_i}{\log \mathcal{L}_{\max} - \log \mathcal{L}_i} \quad (\text{A6})$$

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}_i and $y = 1$ to \mathcal{L}_{\max} , so that

$$P(y) = \frac{d}{2} (1-y)^{\frac{d}{2}-1} \quad [0 < y < 1]. \quad (\text{A7})$$

We now seek the distribution for the maximum likelihood of the live points, $\log \mathcal{L}_{\max}^{\text{live}}$. 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

$$P(y_{\max}^{\text{live}}) = \frac{nd}{2} (1 - y_{\max}^{\text{live}})^{\frac{d}{2}-1} \left(1 - (1 - y_{\max}^{\text{live}})^{\frac{d}{2}}\right)^{n-1} \quad [0 < y_{\max}^{\text{live}} < 1], \quad (\text{A8})$$

which may be roughly summarised as

$$y_{\max}^{\text{live}} \sim 1 - \frac{\Gamma(1 + \frac{2}{d})\Gamma(1+n)}{\Gamma(1 + \frac{2}{d} + n)} \pm \left(\frac{\Gamma(1+n)\Gamma(1 + \frac{4}{d})}{\Gamma(1 + \frac{4}{d} + n)} - \frac{\Gamma(1 + \frac{2}{d})^2\Gamma(1+n)^2}{\Gamma(1 + \frac{2}{d} + n)^2} \right)^{\frac{1}{2}}, \quad (\text{A9})$$

or in the large d, n limit

$$\lim_{d \rightarrow \infty} y_{\max}^{\text{live}} \sim \frac{2H_n}{d} \pm \left(\frac{2(\pi^2 - 6\Psi^{(1)}(1+n))}{3d^2} \right)^{\frac{1}{2}}, \quad (\text{A10})$$

$$\lim_{d, n \rightarrow \infty} y_{\max}^{\text{live}} \sim \frac{2 \log n}{d} \pm \sqrt{\frac{2}{3}} \frac{\pi}{d}, \quad (\text{A11})$$

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

This shows that in general the live points are nowhere near the maximum log-likelihood at any iteration, though they steadily squeeze the interval $[\log \mathcal{L}_i, \log \mathcal{L}_{\max}]$. In

particular, in high dimensions n only gets us harmonically/logarithmically closer, whilst d pushes us linearly further away.

WHCOMMENT: How sure are we of this statement? I agree nested sampling is slower than gradient descent at maximising, but it still converges exponentially. What is the approximation of Eq. (40) in this instance?

WHCOMMENT: Is it worth hiding all these big results in the appendix – it's not that long, and I think readers must read it to pace themselves through the maximum arguments?

WHCOMMENT: $P(\log \mathcal{L})$ for the dead points?

WHCOMMENT: maximum (n) iteration notation?

WHCOMMENT: Fig. 5 move legend.

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
- Baldock R. J. N., Pártay L. B., Bartók A. P., Payne M. C., Csányi G., 2016, *Phys. Rev. B*, 93, 174108
- Baldock R. J. N., Bernstein N., Salerno K. M., Pártay L. B., Csányi G., 2017, *Physical Review E*, 96
- Brewer B. J., Foreman-Mackey D., 2016, *arXiv e-prints*, p. [arXiv:1606.03757](https://arxiv.org/abs/1606.03757)
- Buchner J., 2021, UltraNest – a robust, general purpose Bayesian inference engine ([arXiv:2101.09604](https://arxiv.org/abs/2101.09604))
- Buchner J., 2023, *Statistics Surveys*, 17
- Chopin N., Robert C. P., 2010, *Biometrika*, 97, 741
- Evans M., 2006, World Meeting on Bayesian Statistics Benidorm, 8
- 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
- Gelman A., Meng X.-L., 1998, *Statistical Science*, 13, 163
- Habek M., 2015, *AIP Conference Proceedings*, 1641, 121
- Handley W., 2019, *Journal of Open Source Software*, 4, 1414
- Handley W., 2021, *Physical Review D*, 103
- Handley W., 2023a, Nested sampling: powering next-generation inference and machine learning tools for cosmology, particle physics and beyond, King's College London, https://github.com/williamjameshandley/talks/raw/kcl_2023/will_handley_kcl_2023.pdf
- Handley W., 2023b, The scaling frontier of nested sampling
- 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., 2018a, *Monthly Notices of the Royal Astronomical Society*
- Higson E., Handley W., Hobson M., Lasenby A., 2018b, *Bayesian Analysis*, 13, 873
- Higson E., Handley W., Hobson M., Lasenby A., 2019, *Statistics and Computing*, 29, 891–913
- Keeton C. R., 2011, *Monthly Notices of the Royal Astronomical Society*, 414, 1418
- Kirkpatrick S., Gelatt C. D., Vecchi M. P., 1983, *Science*, 220, 671
- McEwen J. D., Liaudat T. I., Price M. A., Cai X., Pereyra M., 2023, *arXiv e-prints*, p. [arXiv:2307.00056](https://arxiv.org/abs/2307.00056)
- 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](https://arxiv.org/abs/0906.3544))
- Pártay L. B., Csányi G., Bernstein N., 2021, *The European Physical Journal B*, 94
- Sivia D. S., Skilling J., 2006, *Data Analysis: A bayesian tutorial* (Oxford Science Publications). Oxford University Press
- Skilling J., 2006, *Bayesian Analysis*, 1
- Speagle J. S., 2020, *Monthly Notices of the Royal Astronomical Society*, 493, 3132
- Swendsen R. H., Wang J.-S., 1986, *Phys. Rev. Lett.*, 57, 2607
- Trotta R., Feroz F., Hobson M., Roszkowski L., de Austri R. R., 2008, *Journal of High Energy Physics*, 2008, 024
- Wang F., Landau D. P., 2001, *Phys. Rev. Lett.*, 86, 2050
- Williams M. J., Veitch J., Messenger C., 2021, *Physical Review D*, 103, 103006