Approximating the end of nested sampling

Zixiao Hu, Artyom Baryshnikov, Will Handley

7 August 2023

ABSTRACT

This paper develops a technique to estimate the runtime of a nested sampling run at an intermediate iteration for any nested sampler. The known likelihood at an iteration is fitted, then extrapolated using a Gaussian model likelihood, 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

Users would like a order of magnitude runtime estimate Figure of PolyChord output

2 BACKGROUND

2.1 Notation

We shall begin with a brief description of nested sampling 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_{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

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

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

$$p(t_i) = n_{\text{live}} t_i^{n_{\text{live}}-1}, \quad \mathrm{E}[\log t_i] = -\frac{1}{n_{\text{live}}}, \quad \mathrm{Var}[\log t_i] = \frac{1}{n_{\text{live}}^2}. \tag{5}$$

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 $\boldsymbol{t}$.

3 CONVERGENCE OF A NESTED SAMPLING RUN

Let us now take a top-down view of a nested sampling run as it progresses to observe how it converges.

3.1 Posterior convergence

We will examine how our set of samples evolve over the course of a nested sampling run. The conventional perspective for the progress of a run is with $\log X$ along the x-axis; the prior volume is compressed at a steady rate, only picking up the bulk of the integral/posterior at around $\log X = \mathcal{D}_{\mathrm{KL}}$.

At an intermediate iteration, we have an accurate reconstruction of the posterior only down to the current prior volume $\log X_*$. The live points contain information below this volume, but not much more. 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 $\log X_{\text{min}}^{\text{live}}$ has mean and variance

$$\mathrm{E}[\log X_{\text{min}}^{\text{live}}] = \mathrm{E}[\log X_*] - \sum_{k=1}^{n_{\text{live}}} \frac{1}{n_k} \approx -\frac{i_*}{n_{\text{live}}} - \log n_{\text{live}} - \gamma, \tag{6}$$

$$\mathrm{Var}[\log X_{\text{min}}^{\text{live}}] = \mathrm{Var}[\log X_*] + \sum_{k=1}^{n_{\text{live}}} \frac{1}{n_k^2} \approx \frac{i}{n_{\text{live}}^2} + \frac{\pi^2}{6}, \tag{7}$$

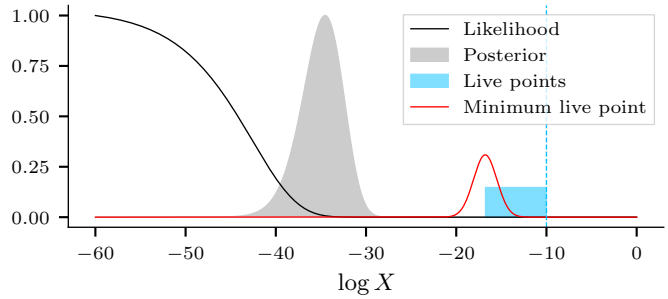

Figure 1.

where the large $n_{\rm live}$ limit is taken for the approximation and γ is the Euler-Mascheroni constant.

This shows that the live points only get us a factor of $\log n_{\rm live}$ closer to the posterior bulk, or 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 begin to resemble the posterior. Intuitively, if the likelihood is sharply peaked then for most of the run the live points are too diffuse to land there with any significant probability.

3.2 Loglikelihood convergence

We now offer an alternate view of the convergence of a nested sampling run, via the loglikelihood instead of the prior volumes. To get an insight into the analytics, work with a Gaussian likelihood of d dimensions and a lengthscale of σ;

$$\log \mathcal{L} = \log \mathcal{L}_{\rm max} - \frac{X^{2/d}}{2\sigma^2}. \tag{8}$$

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

$$P(\log \mathcal{L}) = \frac{1}{\Gamma(\frac{d}{2})} e^{\log \mathcal{L} - \log \mathcal{L}_{\rm max}} (\log \mathcal{L}_{\rm max} - \log \mathcal{L})^{\frac{d}{2}-1} \tag{9}$$

i.e. $2(\log \mathcal{L}_{\rm max} - \log \mathcal{L}) \sim \chi^2_d$, which is the same distribution as that of the dead points up to the latest dead likelihood $\log \mathcal{L}_*$. Meanwhile, the live points are uniformly distributed over the constrained prior and hence have probability distribution

$$P(\log \mathcal{L}) = \frac{d}{2} \frac{(\log \mathcal{L}_{\rm max} - \log \mathcal{L})^{\frac{d}{2}-1}}{(\log \mathcal{L}_{\rm max} - \log \mathcal{L}_*)^{\frac{d}{2}}} \quad [\log \mathcal{L}_* < \log \mathcal{L} < \log \mathcal{L}_{\rm max}], \tag{10}$$

It is helpful at this stage to define a parameter

$$y = \frac{\log \mathcal{L} - \log \mathcal{L}_*}{\log \mathcal{L}_{\rm max} - \log \mathcal{L}_*} \tag{11}$$

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

$$P(y) = \frac{d}{2}(1-y)^{\frac{d}{2}-1} \quad [0 < y < 1]. \tag{12}$$

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

$$P(y^{\rm live}_{\rm max}) = \frac{nd}{2}(1-y^{\rm live}_{\rm max})^{\frac{d}{2}-1} \left(1-(1-y^{\rm live}_{\rm max})^{\frac{d}{2}}\right)^{n-1} [0 < y^{\rm live}_{\rm max} < 1], \tag{13}$$

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

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

This shows that in general the live points are nowhere near the maximum loglikelihood at any iteration, though they do steadily squeeze the interval $[\log \mathcal{L}_*, \log \mathcal{L}_{\rm max}]$. 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:

$$f = \frac{\int_0^{X_{\rm end}} \mathcal{L} dX}{\int_0^{\infty} \mathcal{L} dX}. \tag{15}$$

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. (8) we find the answer in terms of lower incomplete gamma functions

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

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

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

We thus have an expression relating $\mathcal{L}_{\rm end}$ 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:

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

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.

Observation 3: nested sampling as a maximiser

Putting this together, taking $\mathcal{L}_*$ in eq. (11) to be $\mathcal{L}_{\rm end}$, and combining this with eq. (14) we find

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

showing that in general nested sampling will finish at a contour $d/2e$ away from the maximum loglikelihood, 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.

From above analysis we learn that

- Knowledge of the posterior early on in the run is limited by the maximum live point, which is far from the posterior bulk
- At termination, the maximum live point is far from the true maximum –> nested sampling alone is a poor optimiser (but great for finding starting point of gradient descent)

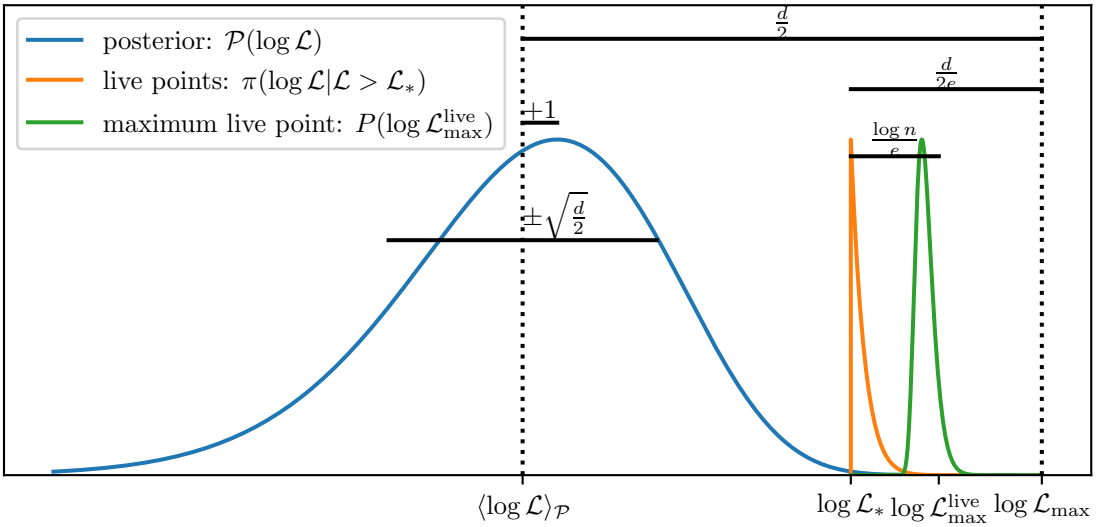

Figure 2.

4 METHOD

Let us now tackle the problem of estimating how many steps we still have to take in a nested sampling run before we can terminate.

The above analysis shows that our knowledge of the posterior at an intermediate iteration is limited by the maximum 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.

For practical purposes, one would never use this approximate posterior to do inference. 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: fitting a model function $f(X, \boldsymbol{\theta})$ to the known likelihood profile allows us to express the prior volume we need to compress to as

$$\Delta \mathcal{Z} = \epsilon \mathcal{Z}_{\rm tot}, \tag{20}$$

$$\int_0^{X_f} f(X, \boldsymbol{\theta})\, dX = \epsilon \left(\int_0^{X_i} f(X, \boldsymbol{\theta})\, dX + Z_{\rm dead} \right), \tag{21}$$

where X_i is the volume of the iteration we have currently compressed to, and $\mathcal{Z}_{\rm dead}$ 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. If $n_{\rm live}$ is constant then at each iteration $\log X$ decreases by $1/n_{\rm live}$, so the total number of iterations N_f will be

$$N_f = -n_{\rm live} \log X_f. \tag{22}$$

One has freedom to choose the form of the fitting function $f(X, \boldsymbol{\theta})$. Following the previous analysis, we use a Gaussian profile, which has the advantage that the regression itself can be sped up analytically, as well as the fact that many real likelihoods are Gaussian or near-Gaussian. However, any other function that is sufficiently flexible to capture the shape of the likelihood profile will do.

Fit a Gaussian profile, with advantages that a) Gaussian error common b) analytic optimisation Extraction of termination logX

Prediction uncertainty dominated by correlations, not least squares i.e. scatter of individual points - get by simulating X

5 RESULTS

Toy examples - pure Gaussian, wedding cake, Cauchy Real cosmological examples - Planck