Nested sampling for physical scientists

Greg Ashton1,2, Noam Bernstein3, Johannes Buchner4, Xi Chen5, Gábor Csányi6, Andrew Fowlie7,8, Farhan Feroz9, Matthew Griffiths9, Will Handley10,11, Michael Habeck12, Edward Higson13, Michael Hobson11, Anthony Lasenby5,10,11, David Parkinson14, Livia B. Pártay15, Matthew Pitkin16, Doris Schneider17, Joshua S. Speagle18,19,20, Leah South21, John Veitch22, Philipp Wacker17, David J. Wales15,23 and David Yallup10,11

Abstract | This Primer examines Skilling’s nested sampling algorithm for Bayesian inference and, more broadly, multidimensional integration. The principles of nested sampling are summarized and recent developments using efficient nested sampling algorithms in high dimensions surveyed, including methods for sampling from the constrained prior. Different ways of applying nested sampling are outlined, with detailed examples from three scientific fields: cosmology, gravitational-wave astronomy and materials science. Finally, the Primer includes recommendations for best practices and a discussion of potential limitations and optimizations of nested sampling.

In 2004, Skilling introduced the nested sampling algorithm1–3 in the context of Bayesian inference and computation (Box 1). The nested sampling algorithm solves otherwise challenging, high-dimensional integrals by evolving a collection of live points through parameter space. The algorithm was immediately adopted in cosmology because it partially overcomes three of the major difficulties in Markov chain Monte Carlo4,5: the algorithm traditionally used for Bayesian computation. Nested sampling simultaneously returns results for model comparison and parameter inference; successfully solves multimodal problems; and is naturally self-tuning, allowing its immediate application to new challenges. Over the past 15 years, research into nested sampling has partially clarified the theoretical properties of the algorithm and its connection to other computational methods. With the development of efficient implementations, variants and cross-checks, the range of nested sampling applications now extends beyond cosmology into many other branches of science.

This Primer summarizes nested sampling origins, principles and algorithm variants. Recent developments are discussed, alongside example results from applications including cosmology, gravitational-wave astronomy, particle physics and materials science. Best practices when using nested sampling and issues with the technique are outlined. The Primer ends with a look to the future of nested sampling and Bayesian computation. Further details are presented in Supplementary information, including a simple numerical example in part F.

Multidimensional integrals
Nested sampling is primarily an algorithm for integration. A general multidimensional integral of a function \( L \) over parameters \( \Theta \) may be written as:

\[
Z = \int L(\Theta)\,d\mu(\Theta)
\]  

(1)

There is often a need to integrate in high dimensions and with challenging integrands. The integrand is assumed to be positive, \( L(\Theta) \geq 0 \).

Often, \( Z \) is a physical quantity, such as the total mass of an object distributed with density \( L \) across volumes \( d\mu(\Theta) \). Although nested sampling is a general method for integration, for concreteness all applications are viewed through the lens of Bayesian inference (Box 1), where \( d\mu(\Theta) \equiv \pi(\Theta)\,d\Theta \) is an element of prior probability with \( \pi \) the prior, normalized by its nature to \( \int \pi(\Theta)\,d\Theta = 1 \). The integrand, \( L \), is the modulating likelihood function and \( Z \) is the evidence. In scientific inference problems, the integral could be over tens or hundreds of parameters, which are required to model fundamental effects, the calibration and systematics of detailed experimental measurements.

The integrand may be rewritten through the elementary factorization known as Bayes’ theorem:

\[
L(\Theta) \times \pi(\Theta) = Z \times P(\Theta)
\]

(2)

where

\[
P(\Theta) = \frac{L(\Theta)\pi(\Theta)}{Z}
\]

(3)
is the posterior normalized to $\int P(\Theta) d\Theta = 1$. Notation apart, all that is done here is to decompose the integrand into a magnitude $Z$ and a shape $P(\Theta)$.

Historically, Bayesian computation focused only on the shape, $P(\Theta)$, partly owing to controversies around Bayesian model comparison, such as its sensitivity to the choices of prior, and partly owing to computational difficulties. However, both shapes and magnitudes matter, especially in the general setting of multidimensional integration beyond Bayesian model comparison. Nested sampling surmounts the challenge by computing shapes and magnitudes simultaneously.

Simplifying multidimensional integrals

Before introducing nested sampling, the general integral in Eq. 1 should be simplified. In traditional Riemann-style integration, the space is decomposed into volume elements $d\Theta$ — typically small cubes — and a sum is performed over them. Small cubes, however, rapidly become infeasible in multidimensional integration because their cost grows exponentially with dimension. This is the curse of dimensionality.

However, it is not necessary to decompose the space into small cubes; the cells can be any shape desired. The integrals needed for quantification are:

$$Z = \int L(\Theta) \pi(\Theta) d\Theta = \lim_{|\Delta\Theta| \to 0} \sum L(\Theta) \pi(\Theta) \Delta\Theta$$  \hspace{1cm} (4)

and so may be defined as limiting sums over volume elements that should be small enough to keep $P(\Theta)$ almost constant regardless of shape. It is, therefore, possible to combine the cells where the integrand is almost constant. Schematically, this can be written as:

$$Z = \sum L(X) \Delta X$$  \hspace{1cm} (5)

where $\Delta X$ is the volume of cells that share likelihood $L(X)$ weighted by the prior $\pi(\Theta)$. This is illustrated schematically in Fig. 1a and works whether the integrand is unimodal or multimodal.

It is possible to reach Eq. 5 more concretely by noting that the evidence is the expectation of a non-negative random variable, such that it may be written as:

$$Z = \int X(L) dL$$  \hspace{1cm} (6)

where the volume variable $X$:

$$X(L^*) = \int_{L > L^*} \pi(\Theta) d\Theta$$  \hspace{1cm} (7)

is the volume enclosed by contour, $L^*$. This result can be readily proven by integration by parts (see also section 21 in Ref. 3). Applying integration by parts again to Eq. 6 obtains the nested sampling evidence identity.

$$Z = \int_0^1 L(X) dX$$  \hspace{1cm} (8)

providing that $L(X)$ — the inverse of $X(L)$ — exists and that the evidence is finite. This formalizes the schematic Eq. 5. This result is more formally discussed in Box 2.

Nested sampling

As the multidimensional integral in Eq. 7 is impractical in high dimensions, a statistical estimation is inevitable. Nested sampling starts with an ensemble of $n_{\text{live}}$ random locations $\Theta$ drawn from the prior, $\pi(\Theta)$, each of which has its likelihood $L(\Theta)$, which can be placed in ascending order. Crudely, if the lowest half of the values are discarded, the survivors would be random samples taken within the restricted volume $L > \text{Median}[L]$, which would statistically be roughly half the original volume. This enables a statistical estimate of the volume variable in Eq. 7. Repeating that $n_{\text{iter}}$ times would yield compression by a factor of about $2^{n_{\text{iter}}}$. This is the exponential behaviour required to overcome the curse of dimensionality.

Nested sampling works by statistical estimates of the compression, a general and fundamental operation that is used in various ways not limited to those in Table 1. The evidence identity in Eq. 6 is not required in every application and many applications only use the compression in Eq. 7.

Formulation

The nested sampling algorithm is presented here in more detail. It is assumed that there are no regions of constant likelihood resulting in likelihood plateaus. The nested sampling algorithm begins by drawing an ensemble of $n_{\text{live}}$ samples from the prior. The likelihood for each sample is computed. The smallest likelihood is denoted by $L^*$ and discarded. The remaining live points are now distributed over a compressed volume; the factor by which the volume is compressed is denoted by $t$. Finally,
Constrained prior
The prior for the parameters restricted to the region in which the likelihood exceeds a threshold.

Bulk
A region with size of order e^n, where N is the Kullback–Leibler divergence, that contains the overwhelming majority of the posterior mass. Closely related to typical sets. Usually, the bulk will not lie near the mode of the posterior, especially in high dimensions.

Tractable
An adjective used to describe problems that are feasible to solve under computational, monetary or time constraints.

Markov chain
A sequence of random states for which the probability of a state depends only on the previous state.

**Box 1 | Bayesian inference**

Although nested sampling is a general-purpose algorithm for integration, its primary application has been for integrals in Bayesian inference. Consequently, the Primer describes nested sampling using that language. In Bayesian inference, the current state of knowledge is quantified by probability, and to learn from data, probabilities are updated using Bayes’ theorem:

\[
P(\Theta) = \frac{P(D|\Theta)P(\Theta)}{Z}
\]

where \(P(\Theta)\) is the prior, \(P(D|\Theta)\) is the likelihood given \(\Theta\), and \(Z\) is the normalizing constant.

A single iteration of nested sampling is shown in Fig. 1c. When most interested in the magnitude of the evidence, \(\log Z\), consider:

\[
\log t = \langle \log t \rangle = \frac{-1}{n_{\text{live}}} \sum_{t} w_{Li}^* 
\]

as under repeated multiplication \(X_t = t_1 t_2 \ldots t_k\), it is the logarithms that add. For a more complete inference, the compression factors \(t\) may be sampled directly from \(\beta(n_{\text{live}})\).

Having obtained estimates of the volume \(X(L^*)\) at each of \(n_{\text{live}}\) iterations, the evidence may be accumulated via Eq. 6 or Eq. 8, for example by the trapezium rule:

\[
Z = \sum_{i=1}^{n_{\text{live}}} w_{Li}^* 
\]

where the weights are:

\[
w_i = \frac{L(X_{i-1} - X_{i+1})}{2}
\]

The sum in Eq. 13 converges to the desired integral. This is the magnitude. The shape is obtained by assigning weights to each sample (see Ref. 3 for discussion):

\[
P_i = \frac{w_i L_i^*}{Z}
\]

normalized such that \(\sum P_i = 1\). These are the posterior weights of the dead points; the shape may be recovered by, for example, a weighted histogram or other density estimation method. The whole algorithm is shown schematically in Box 1 and the summation for a two-dimensional problem in Fig. 1b.

Compared with the description of nested sampling, a single live point per iteration is replaced, rather than half of the live points. Although the number of replacements can be varied, one replacement is optimal because for \(r \ll n_{\text{live}}\), replacing \(r\) points per iteration would reach the posterior bulk in about \(r\) times fewer iterations. Although the computational expense would not change as \(r\) replacements are required per iteration, the error estimates would scale as \(\sqrt{r}\) because reducing the number of iterations increases the relative Poisson noise in the number of iterations.

In the past decade, analogues between nested sampling, statistical mechanics and other statistical methods — including sequential Monte Carlo (SMC) –, subset simulation in rare-event sampling – and more – have been recognized. The connections to SMC and an SMC variant of nested sampling are discussed in Box 4. There are other strategies for computing the evidence; see Refs. 21–23 for reviews. Notable examples include approximating the integrand by a tractable function, density estimation and Gibbs’ sampling, importance sampling and techniques that reuse Markov chain
Monte Carlo draws\textsuperscript{29}. Nested sampling lies in a class of algorithms that form a path of bridging distributions and evolves samples along that path\textsuperscript{30,31}. Nested sampling stands out because the path is automatic and smooth — compression along $\log X$ by, on average, $n^{-1/2}$ at each iteration — and because along the path is compressed through constrained priors, rather than from the prior to the posterior. This was a motivation for nested sampling as it avoids phase transitions — abrupt changes in the bridging distributions — that cause problems for other methods, including path samplers, such as annealing.

**Uncertainties**

Estimates of the magnitude and shape in Eq. 1 should be accompanied by a discussion of the bias and statistical uncertainty. The latter originates from noisy estimates of the compression factors. The resulting statistical uncertainty in the evidence may be estimated by considering the compression required to reach the bulk of the posterior. This may be quantified by the information content\textsuperscript{32,33}:

$\int H = \int \frac{P(\Theta)}{\pi(\Theta)} \log \frac{P(\Theta)}{\pi(\Theta)} d\Theta \quad (16)$

known in statistics as the Kullback–Leibler divergence. This can be written using the volume variable as:

$H = \int P(X) \log P(X) dX$ \quad (17)$

$H = -\int P(X) \log X dX + \int P(\log X) \log P(\log X) d(\log X)$ \quad (18)$

where $P(X) \equiv L(X)/Z$ is the posterior density of the volume. In Eq. 18, the Kullback–Leibler divergence equals minus the posterior expectation of $\log X$ minus the differential entropy associated with the posterior of $\log X$. As the first term typically dominates, the Kullback–Leibler divergence provides a measure of compression.

Consequently, from Eq. 12 it is likely to take about:

$n_H \approx n_{\text{live}} H$ \quad (19)$

iterations to compress to the bulk of the posterior at $\log X = -H$, and this count is likely to be subject to $\sqrt{n_H}$ Poisson variability. Neglecting contributions from outside the bulk, we may write the evidence sum in Eq. 13 as:

$Z \approx e^{-n_H/n_{\text{live}}} \sum_{i=n_H}^{n_{\text{live}}} (1-t)^{-n_H} L_i^*$ \quad (20)$

The final estimate of $\log Z$ will be plausibly subject to an approximately Gaussian uncertainty from the first factor.
As is usual for statistical uncertainties, the above is the transformation that allows a shift of the integration from \([0,1]\). Now, it is possible to define:

\[ \Theta = \text{an overloaded function as we consider separate functions } \Theta(\theta) \text{ and } \Theta(\lambda). \]

**Parameter domain**

The set of a priori possible parameters, usually the reals \(R^n\) or a subset thereof.

**Survival function**

A function \(F(x)\) associated with a distribution that returns the probability of obtaining a sample greater than \(x\).

**Super-level sets**

A \(\lambda\)-super-level set of any function contains all points for which the function value exceeds \(\lambda\).

**Kullback–Leibler divergence**

A measure of difference between two distributions that may be interpreted as the information gained by switching from one to the other.

\[ \Delta \log Z = \frac{H}{n_{\text{live}}} \]

Therefore, nested sampling statistical uncertainties scale as \(1/n_{\text{live}}\) as is usual for statistical uncertainties (see Ref. 7 for an alternative proof and discussion).

Although nested sampling was first introduced with this estimate\(^1\), it can be unreliable. Authority rests with repeated simulation through Eq. 10 of what the compressions might have been. See Keeton\(^34\) for further discussion of the statistical uncertainty in the nested sampling estimates and Supplementary information for discussion of uncertainties in nested sampling estimates of the posterior. As well as this statistical uncertainty, there are four potential sources of bias: bias originating from failure to faithfully sample from the constrained prior; bias from the choice of estimator for the compression factor; the generally negligible quadrature error; and the potentially important truncation error in Eq. 13. The latter occurs due to stopping after a finite number of iterations. Provided that nested sampling is appropriately configured, the statistical uncertainty usually dominates.

Difficulty in nested sampling does not lie in dimension but, rather, in compression from the prior to the posterior. The compression and resolution \(n_{\text{live}}\) alone determine the uncertainty and the runtime. To maintain a given uncertainty, by Eq. 21 \(n_{\text{live}} \propto H\) live points are required and by Eq. 19 \(n_{\text{live}} \propto H^2\) iterations are needed. If the prior and posterior are factorizable into a term for each dimension, by Eq. 16 the Kullback–Leibler divergence is additive, so it scales linearly with dimension, \(D\), meaning runtime goes as \(D^2\). Nested sampling beats the exponential scaling with dimension expected from the curse of dimensionality.

### Experimentation

This section looks at how to implement nested sampling, including considerations when choosing the number of live points, drawing new live points from the constrained prior, parallelization and deciding when to stop.

**Choice of the number of live points**

The number of live points controls the rate of exponential compression during a nested sampling run; compressing inwards by about \(\Delta \log X = 1/n_{\text{live}}\) per iteration. This means that runtime scales as \(O(n_{\text{live}})\) (Eq. 19) and the dominant uncertainties on the evidence integral (Eq. 21) and posterior scale as \(O(1/n_{\text{live}})\). The above considerations are represented graphically in Fig. 2a.

There is, therefore, a trade-off between runtime and uncertainty. Despite that, there is no straightforward method for choosing the number of live points if the required compression is not known ahead of time. Furthermore, the number should be chosen bearing in mind the alternative role that it plays as a resolution parameter for nested sampling\(^35,36\), especially in multimodal problems. In particular, \(n_{\text{live}}\) should be large enough that, at any time, the constrained prior splits into disjoint modes and at least one live point lies inside the footprint of each mode. As a rough rule of thumb, if the constrained prior occupies a total volume \(X\), only modes with a footprint greater than about \(X/n_{\text{live}}\) may be reliably found\(^34-36\). This defines a resolution down to which the posterior can be reliably sampled. Modes with smaller footprints are typically not located and correctly sampled, and will not contribute to the evidence estimate. Moreover, to sample reliably and efficiently from the constrained prior, it is usually advisable that \(n_{\text{live}}\) exceeds the dimensionality of the parameter space.

**Dynamic nested sampling**

So far, this Primer has only considered nested sampling with a fixed number of live points, where uncertainties in both posterior distributions and evidence estimates are reduced by increasing the number of live points. However, the evidence depends on an accurate estimate of the total compression when reaching the posterior bulk. On the other hand, the posterior depends only on an accurate estimate of the relative compression once inside the posterior bulk. The former uncertainty cancels in the posterior weights in Eq. 15, as they are invariant under rescaling of the estimates of the volume variable. This reflects that although the evidence may depend strongly on the size of the prior, the posterior usually depends only weakly on its shape.

Consequently, a dynamic number of live points should be considered to efficiently reduce uncertainties in parameter inference. As shown in Fig. 2a, it is possible to quickly compress to the posterior bulk using few live points. Upon reaching the posterior bulk, the number of live points can be increased, reducing uncertainty in the bulk of the posterior mass. Schemes that vary the
number of live points are denoted as dynamic nested sampling. Open source dynamic nested sampling software packages include dynesty and dyPolyChord (see Table 2). The gains are greatest for problems with substantial compression to the posterior bulk.

In dynamic nested sampling schemes, the number of live points can be automatically adjusted to maximize a user-specified objective for a fixed computational budget. Usually, the run starts with an exploratory nested sampling run with a constant number of live points. The remaining computational budget is spent repeatedly increasing the number of live points in the most important regions of volume, judged according to the objective (for example, the shaded region in Fig. 2b). Because dynamic nested sampling re-works a run and adds extra samples anywhere, running for longer times reduces uncertainties and increases the effective sample size, unlike in ordinary nested sampling.

In the original dynamic nested sampling algorithm, a user specifies their objective by assigning a relative importance to reducing uncertainties in the posterior and the evidence. When focusing on posterior inferences, dynamic nested sampling can achieve orders of magnitude reductions in computational cost for a fixed uncertainty. The approach can also improve evidence calculation accuracy for a fixed number of samples, and simultaneously improve posterior inferences and evidence calculations. This objective was generalized by the reactive nested sampling implementation of dynamic nested sampling. This algorithm considers the computation as a graph, where the nodes are the live and dead points, whereas the edges indicate replacement of one point by another. Multiple agents can then add live points (edges) where needed, and optimize towards additional goals, such as the effective sample size, or the number of samples per cluster. Lastly, a significant variant of nested sampling, diffusive nested sampling, allows the number of live points at a given likelihood threshold to change. In this variant, random walks starting from the existing live points are permitted to step down or up in likelihood, refining the typical likelihood in a volume range $X$.

Table 1 | Applications of nested sampling

| Nested sampling application | Details |
|----------------------------|---------|
| Integration                | Perform the general multidimensional integral (Eq. 1) for positive integrands |
| Global optimization        | Maximize the likelihood, $L$, by compressing to $\Theta$, the maximum of $L(\Theta)$, with no restriction to unimodal distributions; this may require strict settings and be more computationally expensive than integration; see Refs. 135,208 for further discussion |
| Bayesian inference          | Nested sampling simultaneously computes the posterior and the Bayesian evidence, allowing parameter inference and model comparison |
| Approximate Bayesian         | Perform efficient approximate Bayesian computation by applying nested sampling to the joint space of parameters and data |
| Statistical thermodynamics  | If the Boltzmann factor is used as the likelihood, $L(\Theta) = e^{-\beta E(\Theta)}$, where $E$ is the energy of the state $\Theta$ among $N$, nested sampling may be used to compute the partition function; by accumulating for several temperatures $T$ in parallel, thermodynamic functions, such as the specific heat $C_V = \frac{dS}{dT}$, can be plotted as functions of temperature without needing multiple runs |
| Rare-event sampling         | Volume variable $X$ may be interpreted as the probability of a rare event or used to compute a $P$ value in frequentist statistics |

Exploration strategies

The nested sampling algorithm progresses by replacing live points with independent samples drawn from the constrained prior in Eq. 5, namely the prior restricted to regions in which the likelihood exceeds a threshold. This is the main difficulty for efficiently and reliably implementing nested sampling, especially in multimodal problems. As long as the constrained prior is successfully sampled, nested sampling works ideally in unimodal and multimodal settings.

Although it is possible to sample from the entire prior until a sample is found with a likelihood that exceeds the threshold, this rapidly becomes inefficient due to the exponential reduction in the volume contained within the constrained prior at every iteration. Fortunately, the current set of live points and the estimate of the volume enclosed by the contour may guide the search for new live points. There are two main approaches for sampling from the constrained prior: region samplers and step samplers. These are illustrated in Fig. 2b. Analogous to the choice of transition kernels in Markov chain Monte Carlo, the options here lead to various flavours of nested sampling with different performance characteristics and different behaviour as the dimension grows. A priori, they require as much tuning as Markov chain Monte Carlo; however, the live points enable them to build proposal structures and apply clustering algorithms — analogous to ensemble samplers — meaning nested sampling is naturally amenable to robust self-tuning.

As the sampling methods are guided by the live points, their reliability and efficiency usually improve when the number of live points is increased; see Refs. 41,48 for numerical investigations. For both region and step samplers, if no current live points lie inside a mode in a multimodal problem, that region of the constrained prior will not be sampled and the mode is missed.

Typically, both samplers operate in the hypercube, a parameterization where the prior is uniform over a unit hypercube. As a result, the problem of sampling uniformly from within a contour defined by the threshold is slightly simplified. Usually, this is achieved by the inverse transformation method, where a user specifies their

Push-forward measure
(Also known as image measure). The distribution of a random variable under a probability measure.

Transition kernels
Functions that describe the likely steps of a Markov chain.
Box 3 | Schematic of the nested sampling algorithm for the general multidimensional integral

1. Choose an estimate of the compression factor, for example, \( t = e^{-1/\ln 10} \)
2. Initialize volume, \( X = 1 \), and integral, \( Z = 0 \)
3. Sample \( n \) live points from the prior — the live points
4. Repeat
5. Let \( L^* \) be the minimum \( L \) of the live points
6. Replace live point corresponding to \( L^* \) by one drawn from the prior subject to \( L > L^* \)
7. Increment the estimate of the integral, \( Z = Z + L^* \Delta X \), with, for example, \( \Delta X = (1 - t)X \)
8. Contract volume, \( X = DX \)
9. Until stopping criteria satisfied
10. Add estimate of remaining evidence, for example, \( Z = Z + \bar{L} \) \( X \) where \( \bar{L} \) is the average likelihood among the live points
11. Return estimate of integral, \( Z \)

Region sampling. In region sampling, the aim is to construct a region that bounds the iso-likelihood contour defined by the threshold. To find the region, a geometric shape is constructed around the current distribution of live points. The shape must contain at least the currently estimated volume. Then, the independent and identically distributed samples are drawn from within that region until a sample is obtained that passes the current likelihood threshold. To be confident that the region did not encroach on the contour, implementations of region sampling often expand the region by a user-specified factor or by a factor found through cross-validation. For example, by dividing the live points into training and test sets, and ensuring that the region found by the training set includes points in the test set43,46. The expansion factor improves reliability at the expense of efficiency.

The simplest region sampler would involve sampling from the entire unit hypercube, which rapidly becomes inefficient. Instead, most region samplers attempt to estimate the constrained prior by wrapping the live points with one or more overlapping ellipsoids. Using more than one ellipsoid allows complications and multimodal likelihood contours to be efficiently bounded. An appropriate number of ellipsoids can be found by applying clustering algorithms to the live points to identify distinct modes, such as x-means47. The shape and location of the ellipsoids may be approximately found from the mean and covariance of the live points they contain, and a tuning parameter may expand their volumes. This was successfully implemented in MultiNest45. Alternatively, MLFriends45,46 places an ellipsoid around every live point and determines the ellipsoid scale by bootstrapping, a process similar to kernel density estimation with a uniform kernel.

Region samplers face two major limitations when the dimension or complexity of the contour grows. First, accurately bounding a complicated contour depends on the number of live points. For instance, a substructure may not be identified when there are too few points. This results in overly large regions that bound complicated substructures, with poor efficiency. Alternately, as live points are distributed uniformly within the constrained prior rather than near the edge, wrapping a small number of live points can result in overly small estimates of the constrained prior. Second, the accuracy and efficiency of region samplers suffer from the curse of dimensionality. To examine this in more detail, consider region sampling with a single ellipsoid. Suppose that the true contour is a unit hypercube. The smallest ellipsoid that could be constructed to enclose the contour would be a sphere of diameter \( \sqrt{D} \). The volume of this sphere expands exponentially as dimension increases, leading to \( \mathcal{O}(e^{-D}) \) efficiency. This follows the general result that an exponentially increasing fraction of volume lies near the boundaries of high-dimensional shapes. As a result, region samplers are efficient and practical only for moderate to low dimensionality (\( D \ll 20 \)).

Step sampling. Step samplers, by contrast, do not attempt to construct a region that bounds the likelihood contour, avoiding some of the issues of region samplers. Instead, they evolve a randomly chosen existing live point through a sequence of steps to a new approximately independent position. The acceptance rule is simply that we accept a transition to \( \Theta^* \) if:

\[
L(\Theta^*) > L^* \tag{23}
\]

Each step must stay inside the contour. Such step samplers are akin to running constrained Markov chain Monte Carlo inside nested sampling, and were the originally proposed solution. Strategies for generating new positions vary widely, and currently include random-walk Metropolis, where new positions are proposed based on a local target distribution such as a multivariate Gaussian; ensemble proposals, which use the distribution of all live points to propose new positions using strategies such as differential evolution; and slice sampling variants, where new positions within the constrained prior are proposed along a randomly chosen principal axis — slice sampling52,53 — or randomly chosen direction — hit and run54,55. Finally, gradient-based trajectories39,56–59 reflect the current likelihood constraint.

See References41,60,61 for further discussion. In step samplers with a step size parameter, such as random-walk Metropolis, the step size is often tuned to ensure that a substantial fraction (20%) of proposed positions are accepted. This avoids an unacceptable overall sampling efficiency47. This tuning may be performed continuously throughout a nested sampling run47,56,57, although this can introduce biases42.

It can be challenging to judge the number of steps required to ensure that the live points are independent draws from the constrained prior (see Supplementary information). Whereas mild violations of this requirement might be inconsequential47, strong violations lead to unreliable nested sampling evidence estimates42.
The number of iterations required for new samples to approximately de-correlate scales as $O(D^3)$ for random-walk proposals with tuned step sizes, and as $O(D^{1+\alpha})$ for slice sampling or gradient-based trajectories. In practice, the number of steps is often chosen to be $aD^\alpha$, where $a$ is of order one and $b$ is the anticipated dimensional scaling.

The computational cost scales linearly with the number of steps. Unlike region samplers, step samplers escape the curse of dimensionality as their cost shows only polynomial $O(D^b)$ scaling with dimensionality. Nevertheless, region samplers are often more efficient in low dimensions. As a result, step samplers are more often used when applying nested sampling in high dimensions ($D \geq 20$).

Table 2 compares the approaches in several publicly available nested sampling implementations that originated in different research fields. For a multitude of programming languages, there are well-documented, free and open source codes. Support for parallelization to computing clusters and checkpointing are also common features. They usually work with the logarithms of the evidence and likelihood, as the latter may be numerically tiny such that it cannot be represented as a floating point value. In general, sensible defaults for numerous region and step samplers have been found to work across a large variety of problems.

### Parallelization

To make best use of computing resources, computation of the likelihood function could be parallelized. If that is impossible or impractical, an alternative is to design an efficient scheme for parallelizing the nested sampling run itself. As discussed in Refs.13,14, statistically independent nested sampling runs may be combined into an equivalent nested sampling run with $\sum n_{\text{live}}$ live points.

To achieve this, pool the initial live points from the independent runs $a, b, \ldots$ together, giving $\sum n_{\text{live}}$ live points drawn from the prior. Remove the worst, supposing it lay at $L^*$ and originated from run $a$. To draw a replacement from the prior subject to $L > L^*$, simply take the replacement used in run $a$, as it is already a draw from the prior subject to $L > L^*$. This process may continue, weaving together independent nested sampling runs to build a new nested sampling run with $\sum n_{\text{live}}$ live points.

This allows parallelization of a nested sampling run with $n_{\text{live}}$ live points into $n_{\text{CPU}}$ independent runs with about $n_{\text{CPU}}/n_{\text{live}}$ live points each. The independent runs themselves proceed linearly and may be ultimately combined, resulting in a speed-up of about $n_{\text{CPU}}$. Even simpler, estimates of the evidence integrals may be combined by weighted averaging.

However, the reduction in the number of live points per run, $n_{\text{live}}/n_{\text{CPU}}$, impacts the exploration schemes, especially for ellipsoidal rejection sampling, which could lead to inefficient or faulty bounding ellipsoids. It may therefore be desirable to use parallelization within individual nested sampling runs. For example, by drawing $n_{\text{CPU}}$ candidate replacement points and evaluating their likelihoods in parallel. It is thus possible to, subsequently, replace the worst $n_{\text{CPU}}$ live points at each iteration15,16, or to replace a single live point and consider the other evaluated points at subsequent iterations17. An alternative is to replace a single live point and discard as many as $n_{\text{CPU}} - 1$ acceptable candidate points18. The latter is wasteful if more than one viable point is likely to be found among the $n_{\text{CPU}}$ candidates:

$$\text{Speed-up} = \min [n_{\text{CPU}}/1/e]$$  
(24)

if points are considered at subsequent iterations:

$$\text{Speed-up} = n_{\text{live}} \log (1 + n_{\text{CPU}}/n_{\text{live}})$$  
(25)

giving a speed-up of about $n_{\text{CPU}}$ if $n_{\text{CPU}} \ll n_{\text{live}}$. The expression originates from the threshold increasing as the run progresses, meaning that points drawn from the constrained prior might not be valid at a subsequent iteration. Lastly, replacing $n_{\text{CPU}}$ points in parallel per iteration results in a speed-up of about $n_{\text{CPU}}$, but increases the variance in the evidence integral by about $\sqrt{n_{\text{CPU}}}$.

Increasing $n_{\text{CPU}}$ by a factor $\sqrt{n_{\text{CPU}}}$ to maintain the same uncertainty decreases the speed-up to about $\sqrt{n_{\text{CPU}}}$.

### Stopping conditions

An important decision to make is when to stop a nested sampling run. Only a finite number of iterations can be performed, which introduces a truncation error in Eq. 13 that would ideally be negligible. Skilling originally proposed to stop nested sampling after reaching the posterior bulk at $X \approx e^{-H}$ at iteration $n_{\text{iter}} \approx n_{\text{live}}/H$, or using an estimate of the remaining evidence. Popular nested sampling software later adopted the latter. In MultiNest,
this was based on the maximum likelihood found so far, $\max LX$, whereas PolyChord chose the mean likelihood, $\mathbb{E}X$. They stopped once $\Delta Z / Z \leq \text{tol}$, where $\text{tol}$ is a user-specified parameter. Upon deciding to stop, the truncation error in the evidence was corrected either by adding an estimate of the remaining evidence or by killing the live points one by one without replacement and incrementing the evidence in the usual manner until no live points remain. The latter is in keeping with the nested sampling approach. For the former, the remaining evidence may be estimated by $\mathbb{E}X$. See Keeton\textsuperscript{34} for further discussion of the statistical properties of this estimate of the remainder. When nested sampling is used to calculate the partition function of a material system (Box 5), physically motivated stopping conditions can be based on the expected minimum energy — negative log likelihood — or the sampled temperature, which is proportional to the derivative of the limiting energy with respect to nested sampling iteration\textsuperscript{65}.

None of these approaches guarantees that summation has not been terminated too early. There could be a spike of enormous likelihood inside the remaining volume. The computational budget cannot be easily anticipated ahead of time. However, runs that are terminated prematurely may still be used to illustrate what was learned about the posterior before termination. Proposals to construct termination criteria for a fixed computational budget that result in unbiased estimates of the evidence have been suggested\textsuperscript{15}. 

**Results**

Nested sampling results in an estimate of the integral in Eq. 1 and, in the context of Bayesian statistics, a weighted set of draws from the posterior distribution. The error estimate in Eq. 21 depends on the compression and cannot be known ahead of time. If the achieved error is unacceptable, the nested sampling run can be repeated with more live points or combined with a new run. The shape of the posterior can be found from the posterior weights by density estimation. There are dedicated software packages for making publication-quality figures of marginalized posterior densities from weighted samples, including anesthetic\textsuperscript{66}, superplot\textsuperscript{67}, pippi\textsuperscript{68}, dynasty\textsuperscript{39}, getdist\textsuperscript{69}, corner\textsuperscript{70} and pygtc\textsuperscript{71}.
There are a few ways to check the results. First, the nested sampling implementation can be checked, rather than the particular run. To do so, compute the evidence integral in Eq. 1 for problems with known analytic solutions\(^1\). Examples include a multidimensional Gaussian, an egg-box function, the Rosenbrock function\(^2\), Gaussian shells and a mixed Gaussian log \(y\) distribution. Similarly, in some cases, the nested sampling estimates of the volume variable at each iteration, \(X(L^*)\), may be checked against analytic results\(^3\). If discrepancies are found, the implementation is suspect. Alternatively, calculations can be repeated and the distribution of results checked to see whether they are consistent with the results expected for correctly working software. Higson et al.\(^4\) describe procedures for doing this, including tests requiring only two nested sampling runs. These are implemented in nestcheck\(^5\).

Second, the particular nested sampling run of interest may also be checked using a test of the insertion indexes\(^6\). If nested sampling draws new live points independently from the constrained prior, the ranks in likelihood of each new live point compared with the current live points should be uniformly distributed. This test is implemented in the anesthetic\(^7\) nested sampling analysis software, which is compatible with PolyChord and MultiNest, and used on the fly in the nessai\(^8\)\(^,\)\(^9\) nested sampling implementation. If this test fails, but other implementation checks pass, the choices of exploration strategy for the problem may be inadequate. Similarly, Stokes and colleagues\(^10\)\(^,\)\(^11\) discuss testing whether live points are uniformly distributed in the unit hypercube in two-dimensional problems. Last, in the context of parameter inference, posterior samples obtained from different nested sampling implementations or from Markov chain Monte Carlo and nested sampling (see for example Ref. 72) can be compared. Alternatively, a simulation-based calibration can be performed to check the expected properties of the posterior.

Most nested sampling implementations stop once the estimate of the remaining evidence appears negligible. This could omit spikes in likelihood lying inside the remaining unexplored volume. If that is a concern, it may be beneficial to optimize the likelihood using a
The state of all degrees of freedom in a physical system, for example, the microstate of a multi-particle system includes the positions and momenta of all particles.

Microcanonical ensemble
Assigns equal probability to states \( \Theta \) with \( E(\Theta) = \varepsilon \) and zero probability otherwise, such that the energy level \( \varepsilon \) rather than the inverse temperature \( \beta \) characterizes a thermodynamic state.

Applications
This section presents the most established nested sampling applications and highlights the advantages in each case. In addition to these existing applications, nested sampling is beginning to be used in many other areas, including signal processing, phylogeny, systems biology, acoustics, nuclear physics, atomic physics, exoplanet searches and geology.

Cosmology
The rapid spread of Bayesian methods in cosmology in the early 2000s was generated by the growth of data, specifically the new cosmic microwave background (CMB) temperature power spectrum measured by the Wilkinson Microwave Anisotropy Probe (WMAP) and the type Ia supernovae distance measurements. However, these new and powerful cosmological data sets raised questions about the cosmological model. The favoured model of the universe included a mysterious accelerating force, the dark energy, which accounted for 70% of the energy density today. The initial spectrum of density fluctuations in the early universe, which were shown to be Gaussian and adiabatic by the WMAP, indicated a period of accelerated expansion at very early times (about \( 10^{-12} \) s after the Big Bang), known as cosmic inflation. Finally, there was still the question of the universe’s missing mass, which generates the gravitational fields required for cosmic structure, known as dark matter. All three of these phenomena — cold dark matter, dark energy and cosmic inflation — had proposed explanations from the field of high-energy theoretical physics, and these model predictions could be combined with the new wealth of cosmological data to be evaluated with respect to each other, model by model. Consequently, general model selection, and, more specifically, computation of the Bayesian evidence using nested sampling, became a tool of choice.

The simplest cosmic inflation model is driven by a single scalar field, which is a particle physics object similar, but not identical, to the Higgs boson. The behaviour of the scalar field that drives inflation is determined by its potential \( V(\phi) \) (\( \phi \) being the value of the scalar field). The formulation of \( V(\phi) \) — as a function of some set of parameters \( \Theta \) — determines the duration of inflation along with the distribution of anisotropies found in the CMB. Martin et al. performed a detailed Bayesian model comparison between 193 inflationary models using the region sampler MultiNest. They used the cosmological observations, including CMB data from the WMAP mission, to discriminate between alternative models for \( V(\phi) \) and found a slight preference for small field inflation models over large field inflation models.

Another area where nested sampling has been used extensively in cosmology is modelling galaxy clusters. Clusters of galaxies are the most massive gravitationally bound objects in the universe and, therefore, can be used to trace the formation of large-scale structures in the universe. Galaxy clusters can be observed through several methods, including X-ray observations, weak gravitational lensing and exploiting the Sunyaev–Zel’dovich effect. Weak gravitational lensing involves distortion of background galaxies’ images by a large mass lying along the line of sight. Weak lensing enables probing the galaxy cluster’s total mass distribution, including the dark matter. Feroz et al. presented a Bayesian approach using MultiNest to detect galaxy clusters in Sunyaev–Zel’dovich data. This included an estimation of parameters associated with the physical model,
assumed for the cluster and quantification of detection using Bayesian model selection. The approach used the statistics of the CMB anisotropies in the likelihood function. Hurley-Walker et al., presented a joint Bayesian analysis of weak lensing and Sunyaev–Zeldovich observations of several galaxy clusters using nested sampling that allowed estimation of the gas fraction of individual clusters. In modern weak lensing surveys such as DES and KiDS, nested sampling forms a critical part of their parameter estimation, model comparison and tension quantification pipelines (FIG. 3).

The universe’s expansion history can also be measured through distance measurements — such as observations of type Ia supernovae — which can be used as standard candles. The important results from the Supernova Cosmology Project and the High-Z Supernova Search Team presented evidence for the universe’s accelerated expansion, requiring the existence of mysterious dark energy acting against gravity. Cosmological parameters were historically estimated from the observations of type Ia supernovae light curves using a $\chi^2$ approach,

which lacks a rigorous procedure for determination of systematic uncertainties. March et al. introduced a Bayesian hierarchical model to determine cosmological parameters. By using nested sampling, they showed that their principled Bayesian approach delivered tighter statistical constraints on the cosmological parameters more than 90% of the time, reduced statistical bias by a factor of ~2–3 times and had better coverage properties than the standard $\chi^2$ approach.

The measurement of the CMB anisotropies by Planck increased the statistical power over the previous experiment, the WMAP. However, it required more sophisticated modelling of galactic foregrounds and instrumental calibration, introducing $\mathcal{O}(20)$ nuisance parameters. Metropolis–Hastings techniques for parameter estimation accommodated these by exploiting the comparatively fast parameters over cosmological ones. Namely, by caching results from previous calculations, parameters can be changed with negligible computational cost providing the cosmological parameters remain fixed. Region-based samplers such as MultiNest

![Fig. 3](Image)

**Fig. 3** | **Cosmological applications of nested sampling.** (a) Cosmological, non-parametric reconstruction of the power spectrum of primordial cosmological fluctuations. Data measured by cosmic microwave background (CMB) satellites. Reconstructions used a linear spline-based procedure, with $N$ movable knots. Each knot location plus cosmological and nuisance parameters are varied in a full nested sampling fit. The evidence is then used to marginalize over $N$ to produce the final plots. (b–d) Nested sampling in cosmological tension quantification. Bayesian evidence computed by nested sampling can quantify the level of disagreement, hidden by marginalization of high-dimensional parameter spaces. Planck CMB data are in tension with both CMB lensing and BAO in the context of curved cosmologies, meaning they should be combined with caution (part b); only including CMB data, a Bayesian model comparison shows a preference for a closed universe relative to flat ones, despite the Occam penalty associated with the additional parameter (orange bar) (part c); and there is also tension between weak lensing data (DES) and the CMB (Planck) (part d). For further information, see Refs. 116, 125, 219. BAO, Baryon acoustic oscillation; COBE, Cosmic Microwave Background Explorer; DES, Dark Energy Survey; SH0ES, Supernovae and H0 for the Equation of State; WMAP, Wilkinson Microwave Anisotropy Probe. Part a adapted with permission from Ref. 219, APS. Parts b and c adapted with permission from Ref. 125, APS. Part d adapted with permission from Ref. 116, APS.
cannot exploit this hierarchy of parameter speeds and could not reliably navigate the now high-dimensional cosmological and nuisance parameter space. To address this, cosmologists turned to step sampling-based strategies as demonstrated in the PolyChord algorithm\(^{53,120}\) which uses slice sampling. Before this, axial slice sampling was applied to systems biology\(^{326,327}\), alongside this, the ability to exploit a hierarchy of parameter speeds was added. Slice sampling was successfully applied throughout the 2015 Planck inflation paper\(^{36,37}\) to non-parametric reconstructions\(^{25}\) (FIG. 3) and general inflationary model comparison, in particular to the challenging example of axion monodromy models. The non-parametric reconstruction approach — which only became possible with the ability to reliably and fully explore the parameter space of many parameters — was a crucial, model-independent demonstration of the simple power-law behaviour of the primordial power spectrum of density perturbations.

Step sampling was initially introduced in cosmology to exploit a fast–slow hierarchy of parameters. However, the improved dimensionality scaling introduced a new range of cosmological analyses, which were inaccessible to region samplers. Applications include constraining kinetically dominated inflation models\(^{34}\), model comparison for the initial quantum mechanical conditions of the universe\(^{35}\), reconstructions of the dark energy equation of state\(^{35,126,127}\), astronomical sparse reconstruction\(^{38}\), and as a critical role in the GAMBIT combined cosmology and particle analyses\(^{128,129}\). Step sampling takes a leading role in the REACH cm global cosmology analysis\(^{131}\), whereas at the other end of the astrophysical scale it was applied to high-dimensional exoplanet analyses\(^{130,131}\).

**Particle physics**

Particle physics is a field related to cosmology that has also seen various applications of nested sampling. Around 2010, when nested sampling tools such as MultiNest were reaching maturity, the particle physics community focused on the Large Hadron Collider (LHC). A particularly favoured theoretical framework was supersymmetry (SUSY)\(^{32}\). SUSY introduces an array of new particles and unknown parameters that can be fit to collider data and observations from other experiments in a global fit. The goal of a global fit is to understand predictions about future experiments and the best ways to discover new particles.

Global fits of O(10) free parameters in SUSY models presented a problem that nested sampling tools were well suited to solving due to their multimodal nature. A package named SuperBayeS\(^{133–135}\) used the MultiNest implementation of nested sampling and was used to make several early LHC predictions and fits\(^{136–141}\). As results arrived from the LHC, the theoretical landscape shifted. A wider set of models began to be considered using nested sampling\(^{142–145}\). More recently, the GAMBIT collaboration has driven many global fits, using MultiNest and PolyChord to sample parameter spaces, compute Bayesian evidences and benchmark nested sampling against other Monte Carlo-based and gradient-free methods\(^{146,147}\). Related problems, such as tuning phenomenological parameters in event generators, have seen preliminary work and could be rich avenues of nested sampling application in the field. Lastly, nested sampling was recently applied to the sampling space, rather than the parameter space, of a statistical model. This enables efficient computation of small \(P\) values that are used in the discovery of new particles at the LHC\(^{148}\).

**Gravitational waves**

Gravitational-wave astronomy has grown rapidly since the first observation of two colliding black holes\(^{149}\) in 2015 by the LIGO\(^{150}\) and Virgo\(^{151}\) interferometers. The signals are produced by non-axisymmetric changes in the gravitational field, typically sourced by the rapid motion of neutron stars and black holes in binary systems\(^{152}\). The binary orbits decay through gravitational-wave emission, increasing the orbital frequency and the rate of energy loss until the objects merge. Such events, known as compact binary coalescences, produce signals at frequencies of 10–1,000 Hz, which are recorded in the detectors as a time series.

Early development of Bayesian methods for gravitational-wave analysis was contemporary with the first publication of nested sampling. It was found that nested sampling provided an efficient means to robustly sample the posterior distributions of gravitational-wave signals\(^{153}\). This is important as the posteriors offer a rich new astrophysical view: from measuring the masses of neutron stars to the expansion rate of the universe itself. The small signal-to-noise ratios of observed signals and degeneracies in the model parameter space produce posterior distributions that are often multimodal and highly correlated. To date, only nested sampling and Markov chain Monte Carlo approaches have been able to sample the posteriors of all observed events robustly. Although each has its advantages, the agreement between nested sampling and Markov chain Monte Carlo approaches has been critical in building confidence in results. However, the nested sampling approach is better suited to providing robust evidence estimates for model comparisons\(^{154}\). The efficiency of nested sampling does not depend on the use of problem-specific proposals; massively parallel approaches using the dynesty code have made analyses with more advanced signal models\(^{155}\) computationally tractable. The success of nested sampling for analysing merging binary signals inspired many other efforts, such as analysis of continuous signals from individual rapidly rotating non-axisymmetric neutron stars\(^{156,157}\); detection of unmodelled sources\(^{158}\); model selection between different physical mechanisms of core-collapse supernova encoded in the gravitational-wave signal\(^{159}\); and detection of a stochastic superposition of weak merger sources\(^{160}\).

Compact binary coalescences consisting of two black holes are the dominant sources of gravitational-wave signals seen by the LIGO and Virgo detectors\(^{152}\). The signals are modelled through a combination of post-Newtonian approximations to general relativity and relativistic numerical modelling\(^{157,158}\). The signal model, as observed in a detector, is parameterized by eight parameters that...
are intrinsic to the binary system — the individual masses, \( m_1 \) and \( m_2 \), and their three-dimensional spin vectors — and seven parameters related to the relative orientation and position of the system with respect to the detector, including the source’s luminosity distance \( D_L \) and location in the sky. For systems that include at least one neutron star, the signal model’s phase evolution requires additional parameters related to the neutron star equation of state\(^{185}\).

Signals from black hole mergers last only a few seconds within the sensitive frequency regime of the detectors. The detector data can be approximated as noise drawn from a stationary Gaussian process, described by a known power spectral density on these timescales. This approximation can either be estimated from data surrounding the signal\(^{166}\), or inferred directly using a parameterized model of its shape\(^{167}\). For inference, a Gaussian likelihood function of the form given by Whittle\(^{168}\) can be used\(^{166}\). For multiple detectors with independent noise, the likelihoods can be coherently combined using the product rule of probability. The prior distributions used are discussed by Veitch et al.\(^{166}\).

They are generally set to be uninformative, for example uniform over a sphere for sky coordinates, or constrained to be within a physically reasonable range. In addition to the parameters related to the source, the likelihood can also contain \( \sim 60 \) unknown parameters that describe the frequency-dependent uncertainties in the phase and amplitude calibration of the detectors\(^{185}\).

For an observed signal, using the likelihood and priors discussed, the joint posterior of these 15 (or more) parameters can be extracted after the application of nested sampling. The posterior distribution of typical events observed so far is not unimodal or Gaussian. Owing to the relatively small signal-to-noise ratio, they exhibit significant degeneracies and correlations. For example, the posterior of the source sky location is primarily determined by differences in arrival time of the signal at different detectors. This produces ring-shaped degeneracies as demonstrated in Fig. 4b.

The many degeneracies encountered for typical compact binary coalescence inference have led to significant work to identify optimal parameterization and, where possible, use a marginalized likelihood\(^{170}\). An optimal parameterization involves identifying a mapping between the physical model parameters and combinations that reduce the complexity of the target density. Usually, a good re-parameterization involves identifying the combinations of physical parameters which are best measured. As an example, the physical mass of the two-component stellar objects are \( M_1 \) and \( M_2 \). However, there is a strong banana-like correlation between the two masses (Fig. 4a) and an exact degeneracy under exchange of \( M_1 \) and \( M_2 \). To enable efficient sampling, Veitch et al.\(^{166}\) propose sampling in the chirp mass, \( \mathcal{M} \), and mass ratio, \( q \): two algebraic combinations of the component masses. In Fig. 4b, it is shown that the posterior, as viewed in \( M_1 - M_2 \) space, follows contours of the chirp mass. Physically, this is because the chirp mass is the most well-measured mass parameter and has the smallest posterior width, followed by the mass ratio.

The chirp mass and mass ratio parameterization, along with numerous others, have greatly improved sampling efficiencies. The choice of sampling parameters chosen for computational efficiency is separate from the selection of prior distributions. If required, a Jacobian transformation\(^{171}\) can be made to enable sampling in the optimal sampling parameters while setting priors on the physical parameters.

**Materials science**

Nested sampling can be used to study the thermodynamic properties of molecules and materials, which ultimately derive from the partition function:

\[
Z(\beta) = \int e^{-\beta(H(q,p) - q^2)} dq dp
\]

(26)

where \( q \in \mathbb{R}^{3N} \) specifies the spatial coordinates of the \( N \) particles in the system, \( p \in \mathbb{R}^{3N} \) is the corresponding momentum vector, \( H(q,p) \) is the Hamiltonian and \( \beta = 1/k_B T \) is the inverse temperature.

The classical Hamiltonian can be separated into the configuration-dependent potential energy, \( U(q) \), and
momentum-dependent kinetic energy, $K(p) = \sum p^2/2m$ (where $m$ is the mass of each particle), giving:

$$Z(\beta) = \int e^{-\beta K(p)} dp \int e^{-\beta U(q)} dq$$ (27)

For the fully quantum treatment using nested sampling, see Szekeres et al.\textsuperscript{172}. The first factor is a Gaussian that may be computed analytically:

$$Z_\beta \equiv \int e^{-\beta K(p)} dp = \frac{1}{N!} \left( \frac{2\pi m}{\beta \hbar^2} \right)^{3N/2}$$ (28)

and the second, configuration-dependent factor is computed by nested sampling. From the partition function, all thermodynamic quantities of relevance can be calculated, for example the average energy:

$$\langle H(q, p) \rangle = -\frac{\partial \ln Z(\beta)}{\partial \beta}$$ (29)

and the heat capacity:

$$C_v = \frac{\partial \langle H(q, p) \rangle}{\partial T}$$ (30)

These may be interpreted as posterior expectations to be computed after the nested sampling run. In this application, the factor $e^{\beta U(q)}$ plays the role of the likelihood function, parameterized by the inverse temperature, $\beta$; the prior, according to the ergodic hypothesis, is uniform; and the dimensionality is commonly in the order of $10^5$--$10^6$ or more.

The striking difference from standard Bayesian inference is the inverse temperature parameter $\beta$ in the likelihood. Rather than having a single inference problem for some fixed value of $\beta$, almost always, the behaviour of observables as a function of temperature is of interest. Effectively, there is a continuous family of inference problems. For this, nested sampling and other density of state methods (see Box 5) have a remarkable feature: the likelihood is a monotonic function of $\beta$ and, therefore, the entire nested sampling algorithm is invariant to changes in $\beta$. In practice, this means that a single nested sampling run can be used to calculate observables at all temperatures. For large $\beta$ values (corresponding to low temperatures), the partition function is dominated by the lowest energy minima (highest likelihood modes), and for molecules and materials, these configurations correspond to the globally stable structures. On the other hand, when $\beta$ is small (corresponding to high temperatures), the partition function is dominated by the large volume associated with high energy states — or in the parlance of materials science, entropic effects\textsuperscript{173,174}.

Some of the most interesting phenomena in molecular and materials science are associated with this regime change. Collectively known as phase transitions, they are characterized by a dramatic change of where the bulk of the posterior mass lies, as the temperature — or other system parameters, such as pressure — is varied. This makes it very challenging to study phase transitions using numerical sampling schemes. Experimentally, phase transitions are often observed indirectly as changes in the expectation value of observables. For example, discontinuously, in the case of first-order phase transitions such as melting and evaporation, or more directly as sharp peaks in response functions, such as the heat capacity or magnetic susceptibility. Nested sampling promises to enable the calculation of such response functions, in general, with high reliability and minimum fuss.

There are some aspects of the materials application of nested sampling that are favourable compared with the general inference problem. The first is a relatively easy stopping criterion for the nested sampling iterations. For any given model of the potential energy, it is generally not hard to come up with a good global lower bound on the energy, which translates into an upper bound in the likelihood. As the likelihood values sampled by nested sampling appear to converge, if this is close to the known bound, the iterations can be stopped without the risk of missing the highest likelihood mode.

Second, convergence of nested sampling with the number of live points and other sampling parameters is desired and observed in convergence of the heat capacity peak locations. This typically occurs far earlier than the decorrelation of the sampler chains used to explore the constrained prior.

**Thermodynamics of Lennard–Jones clusters.** The Lennard–Jones potential is a simple model for describing the pairwise interactions between atoms, and provides the basis for benchmarking algorithms for modelling materials. In particular, specific-sized clusters of Lennard–Jones atoms exhibit complex thermodynamic properties due to solid–solid transitions — the finite system analogue of first-order phase transitions mentioned previously — caused by the existence of competing low-energy minima, in addition to solid–fluid melting\textsuperscript{175,176}.

Some features of atomistic energy landscapes are illustrated in Fig. 5a using the example of the cluster of 38 particles (LJ\textsubscript{38}) along with the corresponding disconnectivity graph\textsuperscript{175,178}, which help convey the relationships between the large number of local energy minima (likelihood modes). Each leaf of the tree structure corresponds to a local minimum of the energy, or to a closely related set, and junctions represent saddle points connecting the minima. As the number of particles in the system increases, the number of distinct local minima grows exponentially. According to one estimate, LJ\textsubscript{38} for example, has $10^{15}$ distinct minima, not counting permutation isomers\textsuperscript{174}.

Potential energy disconnectivity graphs are good at showing the topology of the energy minima. However, to reflect the volumes’ free energy, disconnectivity graphs are required\textsuperscript{177,180}. Using nested sampling, energy landscape charts\textsuperscript{96} can be generated, and one is shown for LJ\textsubscript{38} in Fig. 5a. The vertical axis is the potential energy, and the curve shows a series of possibly nested basins, whose width at each potential energy level is proportional to the volume of the corresponding prior mode slice. To fit the chart into a single plot, the volumes are scaled by an exponential factor, whose logarithm is shown on the right-hand axis. Each basin, encompassing a range of
closely related configurations, corresponds to a macroscopic state of the system — illustrative structures are shown. The relative volume of the funnel associated with the global minimum (shown in red) and the entropically dominant minimum (in the centre) is 1:15 at the energy level where the two separate at the resolution of this particular nested sampling run. The relative volume reverses to 16:1 at the energy level where the known lowest barrier path connects the two states(181–183).

Through these landscape charts, the challenges of thermodynamic sampling can be best understood. If there are not enough live points — even though a few may make it into the global minimum basin where it splits off — there is a danger of the population of walkers dying out before the global minimum is reached. This extinction can happen even while the different basins are still connected. However, the volume of states that connects them is so small — the path connecting them is narrow — that, in practice, there is no communication between the basins. This phenomenon is generally referred to as broken ergodicity(174).

Although small Lennard–Jones clusters illustrate and enable the development of benchmarking and parameterization of sampling algorithms, nested sampling is not the most efficient method for calculating observables. Specialized bottom-up algorithms, such as basin sampling(174), are significantly more efficient because they start from a low lying minimum, not necessarily the global minimum. They then build a database of all neighbouring minima by a series of perturbations of the particles and subsequent relaxation. Nested sampling performance for Lennard–Jones clusters can be improved either by allowing the sampler to use the pre-generated local minima database, as in the superposition-enhanced nested sampling(184), or by combining a large number of independent nested sampling runs, each with a single live point, as in nested basin sampling(185). Neither of these enhancements makes nested sampling competitive for studying small and moderate-sized particle clusters.

There is one case where nested sampling appears to be an effective tool for studying even the smallest clusters: determining thermodynamically favourable transformation paths between different states at moderately high temperatures, where harmonic transition state theory is no longer applicable(186).

**Phase diagrams of materials.** In the study of condensed phase systems, nested sampling is particularly useful. There are two reasons for this. First, efficient sampling moves are challenging owing to the geometric constraint. Second, as the system size grows, the phase transitions become sharper, meaning that thermal methods become essentially ineffective for a hundred or more particles.

At high energies, all materials are gases. As the energy decreases, they typically condense into a liquid, freeze into a solid and, sometimes, undergo solid-phase structural transitions. A specific heat curve is shown as an example in FIG 5b for the periodic Lennard–Jones system at a single pressure. In the thermodynamic limit, the discontinuity of the potential energy across the phase transition results in a divergence of the heat capacity.
Consider publishing the nested sampling output, allowing further reuse and checks of divergence, settings, for example fewer steps\(^5\), may be adequate when only concerned about parameter inference.

- Report software version numbers and settings.
- Describe priors and likelihood in adequate detail.
- Ideally, publish the code used for the computation, enabling the calculation to be replicated and scrutinized.
- Perform cross-checks on the nested sampling run\(^7\), as implemented in, for example, anesthetic. If practical, consider a simulation-based calibration to check results.
- Report the triplet of log Z, the associated uncertainty and the Kullback–Leibler divergence, H, for each nested sampling computation.
- Consider publishing the nested sampling output, allowing further reuse and checks of the nested sampling run.

However, in a numerical simulation, they are broadened by finite size effects and appear as sharp peaks. Performing the sampling at a range of pressure values, the loci of the peaks define the boundaries between stability regions of different phases in the pressure-temperature phase diagram. There is a structural similarity of the solid phase, a closed packed face-centred cubic crystal, to the narrow global minimum of the LJ\(_{19}\) cluster. As the cluster size grows, the repeated occurrence of these face-centred cubic-like clusters hints at the ground state of the infinite crystal. There is no periodic analogue for the broad, icosahedral basin of the clusters because the fivefold symmetry is incompatible with periodicity.

Nested sampling has successfully characterized the behaviour of a wide range of materials. These include model systems such as the Potts model\(^1\),\(^2\),\(^3\), hard spheres\(^4\),\(^5\),\(^6\),\(^7\), Lennard–Jones clusters\(^8\),\(^9\),\(^10\) and the Jagla potential\(^11\) as well as more chemically realistic potentials for aluminium and the shape memory alloy NiTi\(^12\),\(^13\), lithium\(^14\) and iron\(^15\). Recently, machine-learning interatomic potentials\(^16\) have begun to be applied\(^17\) to increase the predictive power of calculated phase diagrams, leveraging the increased accuracy of the potentials. A detailed overview of materials applications can be found in REF.\(^18\).

**Reproducibility and data deposition**

A set of minimum considerations and reporting standards for nested sampling computations are recommended in BOX 6. The number of live points and stopping conditions should be clearly stated. In addition, implementation-specific settings should be reported; for example, the number of repeats using slice sampling or the enlargement factor using ellipsoidal sampling. If using a public software package, report the version number. As nested sampling is a Monte Carlo algorithm, fixing the random seed is recommended to replicate identical results.

As nested sampling computes an integral, the integrand should be explained clearly, including the choice of likelihood and prior, and any overall constant factors that are sometimes omitted from the likelihood. During publication, computer code may be published alongside the research. Similarly, the nested sampling output files can be publicly deposited. This enables further scrutiny and reuse of the nested sampling results. The output data should be accompanied with sufficient metadata, such as column labels\(^19\).

Although the final result might be a ratio of nested sampling results — for example, a Bayes factor — reporting the results of all individual nested sampling calculations is recommended. This includes the triplet of log Z, the estimated uncertainty and the Kullback–Leibler divergence, H. The first two are most relevant for inference, whereas the Kullback–Leibler divergence indicates the numerical challenge, as it impacts runtime (Eq. 19) and uncertainty (Eq. 21). The effective dimensionality or an Occam factor may also be reported\(^20\).

The nested sampling error estimates are usually reliable. When reproducing nested sampling computations, there should be agreement within uncertainties with the original calculation. However, it is log Z rather than Z alone that is distributed with a roughly symmetric Gaussian error.

**Limitations and optimizations**

**Limitations**

Although nested sampling is broadly applicable, there are several potential limitations. The first limitation relates to the prior. Sampling occurs from the constrained prior and therefore requires a proper prior. For many nested sampling implementations, this proper prior must be transformed from the unit hypercube. Normalizing flows were recently proposed\(^21\) for cases where this is inconvenient or impossible analytically, including using the posterior from a nested sampling run as a prior.

Furthermore, there are limitations related to the likelihood. Integration using nested sampling requires a non-negative integrand, although the compression itself makes no such restriction. Whereas this condition is always fulfilled in statistical applications, it could be violated when nested sampling is used as a general-purpose integrator. Equation 6 was written assuming that \(L \geq 0\), whereas Eq. 8 makes no restriction. Nested sampling, however, compresses upwards in likelihood, such that positive and negative likelihood regions of the integral would be treated differently, with the latter explored at inadequate resolution. In addition, nested sampling requires a tractable likelihood. For cases in which the likelihood is intractable, Mikkelsen and Khammash\(^22\) proposed a likelihood-free nested sampling in the context of systems biology, assuming that an unbiased estimator of the likelihood was available.

Lastly, plateaus in the likelihood function spoil the estimates of the compression\(^2\),\(^2\),\(^2\),\(^2\),\(^2\),\(^2\). That is, sets A with non-negligible prior mass, \(\mu(A) > 0\), and constant likelihood, which means \(L(\Theta) \equiv c\) for all \(\Theta \in A\). The likelihood function can be modified to remove plateaus by adding a negligible unique independent and identically distributed tie-breaking random draw to the likelihood or promoting that draw to a parameter and increasing the dimension of the problem. Alternatively, the algorithm itself may be modified such that it sums plateaus.
correctly but reduces to the standard nested sampling algorithm in their absence. In the presence of plateaus, the minimum likelihood among the live points, \( \min L \), may be shared by several live points.

For example, nested sampling can be modified by removing all \( q \) points in the plateau and then replacing them all\(^{203}\). In this case, the compression factors follow \( \beta(n_{\text{live}} + 1 - q, q) \). This may be applied retrospectively to any nested sampling runs, although it suffers from increased uncertainty as the number of live points is temporarily reduced to as few as \( n_{\text{live}} - q \). Alternatively, the live points may first be increased by sampling subject to \( L > L^* \) until \( n_{\text{live}} - 1 \) lie at \( L > \min L \). Note that \( \min L \) may decrease as live points are added at \( \min L > L > L^* \). Then, \( q \) points are removed in the outermost contour. The compression factor follows \( \beta(n_{\text{live}}, q) \). This cannot be applied retrospectively but the uncertainties are reduced at the cost of greater runtime, as the number of live points is temporarily increased beyond \( n_{\text{live}} \). This may be seen as a dynamic version of the first modification, and the number of live points can be increased according to a different criterion. Lastly, it is possible to split the integral into an integral over plateaus and an integral without plateaus, and perform only the latter with nested sampling\(^{204}\).

Additionally, there are computational limitations. In ordinary nested sampling, running for longer does not reduce uncertainties or increase the number of posterior samples. This is not a drawback when only interested in using nested sampling for optimization. Dynamic nested sampling overcomes this problem by rewinding and resuming an nested sampling run. Further, nested sampling requires \( O(D^2) \) iterations (Eq. 19). The memory required to store the coordinates of every dead point scales as \( O(D^3) \) with potentially challenging scaling factors for clustering in particular exploration strategies. This means that nested sampling implementations that store every dead point become memory bound at \( O(500) \) dimensions.

Users of nested sampling should be aware of ways in which it may fail. Nested sampling may fail to successfully draw independent samples from the constrained prior. Consequently, nested sampling results, including error estimates, may be faulty and the anticipated properties of nested sampling, such as convergence, will not hold. Cross-checks on nested sampling results\(^{27} \) may identify this issue. Owing to this, nested sampling may miss modes. Multimodal problems pose challenges in Bayesian computation, particularly in Markov chain Monte Carlo, where the chain must make a sequence of unlikely steps between modes. Unfortunately, it is impossible to know whether all modes have been found. Broadly speaking, as nested sampling does not depend on slow transitions between modes, it is well suited to multimodal problems. Once a mode is established, it will not be abandoned until the likelihood threshold exceeds that of the points in the mode. Finally, nested sampling may sample inefficiently from the constrained prior. This may occur in high-dimensional problems with rejection sampling strategies.

The checklist presented in BOX 6 includes checks of nested sampling failures.

### Optimizations

There are several ways nested sampling runs may be optimized to make best use of computing resources. First, fast and slow parameters can be used by breaking the likelihood function into fast and slow operations that involve subsets of the \( D \) parameters. The parameters associated with fast and slow operations are referred to as fast and slow parameters, respectively; for example, if the likelihood function is written as:

\[
L(x, y, z, \ldots) = \text{slow}(x) \times \text{fast}(y, z, \ldots)
\]

then \( y, z, \ldots \) are fast parameters and \( x \) is a slow parameter. When selecting a new point, to minimize runtime, the fast parameters should be changed and the slow parameters kept constant, allowing caching of the slow operation\(^{204} \). This approach is natural in the slice sampling exploration strategy, as slices can be picked along fast directions\(^{11} \). If \( D \) parameters are slow, \( D \), rather than \( D \) slow operations are required per iteration. Similarly, this may be exploited in modified Metropolis algorithms\(^{203,205} \) where the proposals change blocks of parameters at a time.

In addition, the posterior can be repartitioned. Nested sampling must compress exponentially through the entire prior volume, which could be slow for a diffuse prior with substantial compression to the posterior. In practice, this limits nested sampling to \( O(100s - 1,000s) \) of dimensions and prohibits the fitting of complex hierarchical Bayesian models and deep neural networks, or reaching Hamiltonian Monte Carlo-like dimensionalities of \( O(10^9) \). In some cases, it may be possible to use a narrower prior and correct the evidence estimates from nested sampling post hoc.

This issue may occur in problems where the prior is unrepresentative\(^{206} \), for example if the observed data lie in the tail of the prior predictive distribution. Bayesian automatic prior repartitioning\(^{207} \) may be used to redefine the prior and likelihood, but leaving their product unchanged. This keeps the evidence integral in Eq. 1 unchanged, but reduces the prior to posterior compression in Eq. 16. For example, a Gaussian likelihood, \( N(x, \sigma^2) \), for a parameter \( x \sim U(-L/2, L/2) \) for \( \sigma \ll L \) has a compression of approximately \( H \approx \log(\sigma/L) \). Repartitioning the prior and likelihood by swapping them obtains \( H = 0 \).

The robustness and efficiency of nested sampling may be improved by exploiting the intrinsic degeneracy between the effective likelihood and prior in the formulation of Bayesian inference problems\(^{111} \). Posterior repartitioning can be viewed as the simplest case — when the importance weight function equals one — of nested importance sampling\(^{7} \).

Rather than performing new nested sampling runs, runs for similar likelihoods and priors can be reused. If the prior or likelihood is modified, it may be possible to reweight the posterior weights and evidence integral\(^{21} \):

\[
P_i' = P_i \times \frac{L_i \pi_i}{Z'} \frac{Z}{L_i \pi_i}
\]
Pseudo-importance sampling

Using algorithms in which an importance sampling density is defined a posteriori.

\[ Z' = Z \times \sum_i P_i \frac{L_i}{N_i} \]  

(33)

where \( P_i \propto w_i \). This reweighting may be interpreted as a pseudo-importance sampling in which the original estimated posterior plays the role of the kernel. This is particularly useful for investigating prior sensitivity in the context of Bayesian inference. The effective sample size of draws from the new posterior may be used to judge the reliability of this procedure. If the new and original posterior distributions differ substantially, a fresh nested sampling run needs to be performed.

There may be cases where it is desirable to investigate several similar likelihood functions at once. Buchner\(^{\text{25}}\) presents a collaborative version of nested sampling that operates on more than one likelihood function at once, where parts of the likelihood evaluation are recycled.

**Outlook**

The Markov chain Monte Carlo computational revolution of the 1990s solved the problem of computing shapes. Skilling’s remarkable nested sampling algorithm solved the problem of computing magnitudes simultaneously. Although it is naturally expressed in the language of Bayesian inference, nested sampling is a powerful and general-purpose integration algorithm. For that reason, nested sampling is expected to remain relevant long into the future. A revolution is currently underway in data science, and high-dimensional spaces and integration are more important than ever.

As reviewed in this Primer, there have been many theoretical developments in understanding nested sampling, including its convergence, errors, diagnostics and techniques for sampling from the constrained prior. These are expected to continue, especially as connections to other statistical methods and machine learning are explored. Theoretical analysis of nested sampling combined with constrained Markov chain Monte Carlo exploration may be aided by the connections to SMC discussed in Box 4. Further developments in understanding uncertainties in nested sampling are anticipated, especially for parameter inference and a dynamic number of live points. Lastly, nested sampling may be considered a meta-algorithm, as it does not specify an algorithm for sampling from the constrained prior. Already, this opening has allowed nested sampling to use developments in machine learning, for example normalizing flows that are beginning to be used to sample from the constrained prior.

The successes and breadth of applications of nested sampling arise from its fundamental simplicity. In the future, nested sampling may be understood and improved in the context of more sophisticated computational methods. However, it will retain advantages and appeal owing to its simplicity.

Published online: 26 May 2022
This work introduces an important dynamic variant of nested sampling that speeds up parameter inference. Brewer, B. J. & Donovan, C. P. Fast Bayesian inference for multimodal problems in cosmology. Mon. Not. R. Astron. Soc. 448, 5206–5214 (2015).

Lavie, B. et al. HeliOS-RETRIEVAL: an open-source nested sampling atmospheric retrieval code applied to the HR 8799 exoplanets and inferred constraints for planet formation. AJ 154, 91 (2017).

Halls, R. D., Thompson, S. J., Handley, W. & Queloz, D. On the feasibility of intense radial velocity surveys for earth-twin discoverers. Mon. Not. R. Astron. Soc. 479, 2968–2987 (2018).

Kitzmann, D. et al. Helios-r2: a new Bayesian, open-source retrieval model for brown dwarfs and exoplanet atmospheres. ApJ 890, 4 (2020).

Aber, E. et al. The Heterogeneous Multi-layered Planetary Atmospheres (HMPA) for southern extra-solar planets — XV. Two Neptune mass planets orbiting HD 13808: A study of stellar activity modelling’s impact on exoplanet lifetimes. Mon. Not. R. Astron. Soc. 503, 1248–1263 (2021).

Elshiekh, A. H., Wheeler, M. F. & Hotil, I. Nested sampling algorithm for the robust model selection, uncertainty quantification, and nonlinear calibration. Water Resour. Res. 49, 8383–8399 (2013).

Trotta, R. Bayes in the sky: Bayesian inference and model selection in cosmology. Phys. Rep. 75, 91–104 (2008).

Spiegel, D. N. et al. First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: determination of cosmological parameters. Astrophys. J. Suppl. Ser. 148, 175–196 (2003).

Reiss, A. G. et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. Astron. J. 116, 1009–1053 (1998).

Perlmutter, S. et al. Measurements of z and I from 42 high-redshift supernovae. ApJ 517, 565–586 (1999).

Guth, A. H. Inflationary universe: a possible solution to the horizon and flatness problems. Phys. Rev. D 23, 347–356 (1981).

Liddle, A. R., Parsons, P. & Barrow, J. D. Formalizing the slow-roll approximation in inflation. Phys. Rev. D 61, 043526 (2000).

Martin, J., Ringeval, C. & Trotta, R. Hunting down the slow-roll approximation in inflation. Phys. Rev. D 83, 063524 (2011).

Allen, S. W., Schmidt, R. W. & Fabian, A. C. Cosmological constraints from the x-ray gas fraction in relaxed lensing clusters. Mon. Not. R. Astron. Soc. 334, L11–L15 (2002).

Allen, S. W. Evard, A. B. & Mantz, A. B. Cosmological parameters from observations of galaxy clusters. Annu. Rev. Astron. Astrophys. 49, 609–740 (2011).

Ferz, F., Hobson, M. P., Cameron, E. & Pettitt, A. N. Importance nested sampling and the MultiNest algorithm. Phys. Lett. B 659, 206–219 (2008).

Machado, J. et al. Measurement of the charged lepton flavour violating parameter $\delta_{\mu\tau}$ in the slow-roll approximation in inflation. Phys. Rev. D 96, 043526 (2017).

Lewis, A., Higdon, J. & Feroz, F. Testing new evidence ratios. Mon. Not. R. Astron. Soc. 391, 1151–1154 (2008).

Feroz, F., Hobson, M. P. & Bridges, M. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. Mon. Not. R. Astron. Soc. 398, 1601–1604 (2009).

Feroz, F., Hobson, M. P., Cameron, E. & Pettitt, A. N. Importance nested sampling and the MultiNest algorithm. Phys. Lett. B 659, 206–219 (2008).
including a nested sampling implementation, *Astronomy and Computing* 37, 100503 (2021).

217. Buchner, J. UltraNest — a robust, general purpose Bayesian inference engine. *J. Open. Source Softw.* 6, 3001 (2021)

218. Albert, J. G. JAXNS: a high-performance nested sampling package based on JAX. Preprint at https://arxiv.org/abs/2005.08602 (2020).

219. Foxley, A., Handley, W. & Su, L. Nested sampling with plateaus. *Mon. Not. R. Astron. Soc.* 503, 1199–1205 (2021).

220. Lewis, A. Efficient sampling of fast and slow cosmological parameters. *Phys. Rev. D* 87, 103529 (2013).

221. Lewis, A. & Bridle, S. Gravitational waves from a 22-solar-mass binary black hole coalescence. *Phys. Rev. Lett.* 116, 241103 (2016).

222. Wales, D. J. The energy landscape as a unifying theme in molecular science. *Phil. Trans. R. Soc. A* 363, 357–377 (2005).

223. van de Schoot, R. et al. Bayesian statistics and modelling. *Nat. Rev. Methods Primers* 1, 1 (2021).

224. D’Agostini, G. Bayesian Reasoning In Data Analysis: A Critical Introduction (World Scientific, 2005).

225. Gregory, P. Bayesian Logical Data Analysis for the Physical Sciences (Cambridge Univ. Press, 2005).

226. van der Linden, W., Dose, V. & von Toussaint, U. Bayesian Probability Theory: Applications in the Physical Sciences (Cambridge University Press, 2014).

227. Baier, K. Nested: pure Python, MPI-licensed implementation of nested sampling algorithms for evaluating Bayesian evidence. *GitHub* https://github.com/kevinb/nestle (2018).

228. Creutz, M. Microcanonical Monte Carlo simulation. *Phys. Rev. Lett.* 24, 185 (1975).

229. Veitch, J. et al. Johnveitch/cpnest: release 0.10.2. *GitHub* https://github.com/johnveitch/cpnest (2017).

230. Moss, A. Accelerated Bayesian inference using deep learning. *Mon. Not. R. Astron. Soc.* 496, 528–538 (2020).

231. Kester, D. & Mueller, M. BayesiFit: a PYTHON toolbox for Bayesian fitting and evidence calculation with the online version containing supplementary material available at https://doi.org/10.1038/s43586-022-00121-x.