Nested Sampling: A Critical and Comprehensive Theoretical Guide

L. Martino*, F. Llorente⊤,

* Universitá di Catania, Catania (Italy).

⊤ Stony Brook University, New York (USA).

2026

Abstract

The nested sampling (NS) technique has gained widespread attention, particularly in cosmology and astronomy, due to its ability to efficiently explore high-likelihood regions - a feature akin to an implicit likelihood optimization that underlies its success. While the full theoretical derivation of NS is complex and involves several approximations, the central challenge lies in sampling from the likelihood-constrained priors, which is crucial for its performance. This work provides a comprehensive and detailed exposition of NS derivation, clarifying both its theoretical foundations and practical challenges. We provide a thorough description of the NS procedure, emphasizing both its strengths and potential limitations. In doing so, this work seeks to deepen understanding of the method and to foster the development of future enhancements, novel variants, and more efficient implementations across a wide range of scientific applications. Thus, the main contribution of this work is twofold: it serves both as a tutorial for newcomers to the field and as a critical review for experienced practitioners.

Keywords: Nested sampling; importance sampling; marginal likelihood; Bayesian inference; MCMC.

1 Introduction

Nested Sampling (NS) [30, 31] is a stochastic quadrature method designed to approximate high-dimensional and potentially complex integrals [20, 24]. The main framework of application is Bayesian inference [13, 20, 29]. Together with the families of importance sampling (IS) and Markov chain Monte Carlo (MCMC) schemes, NS has become a central tool in modern computational statistics and its scientific applications [4, 9]. Excellent, comprehensive and up-to-date reviews of the methodology and its developments can be found in [1, 4]. IS techniques and their sequential versions (also known as particle filters) are widely used in engineering applications, while MCMC algorithms have long dominated the field of statistics. NS, on the other hand, has achieved particular prominence in the physical sciences, especially in cosmology and astronomy [1, 4, 26]. Like IS schemes - and in contrast to standard MCMC algorithms - NS is able to simultaneously provide estimates of model parameters and an approximation of the marginal likelihood [19, 21, 34]. Indeed, originally NS was introduced precisely for the purpose of computing the marginal likelihood (a.k.a., the Bayesian evidence) [30]. Over the past two decades, substantial progress has been made in understanding the theoretical foundations of the algorithm and in elucidating its relationships with other computational methodologies [9]. During this time, numerous efficient implementations, methodological refinements, and diagnostic tools have been proposed to enhance its reliability and performance. As a result, the range of applications of NS has expanded well beyond its original applications in cosmology, finding use across a broad spectrum of scientific disciplines.

This work aims to provide a comprehensive understanding of the NS procedure for all the possible

readers, highlighting both its strengths and potential critical issues. Indeed, although the success of NS is undeniable and widespread, we find it somewhat remarkable (and surprising) when one carefully examines the full derivation of the method. In particular, several authors have pointed out that the derivation of the method relies on a number of approximations, which can affect the statistical properties of the resulting estimators [7, 8, 9, 16]. Motivated by these observations, this work not only presents the methodology in a pedagogical manner but also examines its underlying assumptions and discusses the principal criticisms and theoretical challenges that have emerged in the statistical literature.

The overall NS procedure contains an implicit optimization of the likelihood function that is, in our opinion, the key of the NS success. However, NS relies on the sampling of *likelihood-constrained prior densities*, denoted as $g(\boldsymbol{\theta}|\lambda)$. In our view, sampling from $g(\boldsymbol{\theta}|\lambda)$ can be even more challenging than sampling directly from the posterior. To illustrate this point, note that determining the support of these truncated priors would, in principle, require inverting the likelihood function, that is a task that is generally infeasible or computationally prohibitive. This issue has been emphasized by prominent authors in computational statistics [7, 8], who note that *"in high-dimensional spaces, simulating from the prior until the constraint is satisfied is unrealistic"*. The same authors also discuss the potential poor performance when using vague priors, and the complete impracticality of the method in the case of improper priors [7, 8]. The problem of sampling within domains constrained by likelihood values is, in principle, the same theoretical challenge encountered in *slice sampling* [27]. In both cases, NS and slice sampling, the development of robust and efficient computational implementations, often incorporating sophisticated internal Monte Carlo techniques to draw from $g(\mathbf{x}|\lambda)$, appears to have effectively "solved" the problem, at least from a practical point of view. Another insightful theoretical analysis of NS is presented in [16]. The authors identify and investigate two principal sources of error in NS, both arising from the approximations underlying the derivation of the method. These issues highlight some of the inherent limitations of NS and are closely related to several aspects that we seek to emphasize

throughout this work. A complete advanced theoretical treatments of nested sampling from a statistical standpoint are provided in [9, 16, 18].

Thus, the main contribution of this work is twofold. First, it is intended as a tutorial for readers who are new to nested sampling, providing a self-contained introduction with all details of its derivation. Second, it offers a critical review aimed at experienced practitioners, highlighting both the strengths and the limitations of the approach.

The remainder of the paper is organized as follows. Section 2 introduces the necessary background material and establishes the main notation. The theoretical framework required for a rigorous description of the NS procedure is developed in Sections 3 and 4. A comprehensive and detailed presentation of the Nested Sampling (NS) algorithm is provided in Section 5. Its connections with importance sampling (IS), as well as more advanced NS variants, are examined in Sections 6 and 8. Finally, concluding remarks are presented in Section 9.

2 Background and main notation

In many applications, the goal is to make inference about a variable of interest, $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_{D_\theta}] \in \Theta \subseteq \mathbb{R}^{D_\theta}$, where $\theta_d \in \mathbb{R}$ for all $d = 1, \ldots, D_\theta$, given a set of observed measurements, $\mathbf{y} = \left[y_1, \ldots, y_{D_y}\right] \in \mathbb{R}^{D_y}$. In a standard Bayesian framework, we assume to known an observation statistical model that induces a likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta})$. Assuming a prior probability density function (pdf) $g(\boldsymbol{\theta})$ over the parameter vector to infer, all the statistical information is summarized by the posterior density, i.e.,

$$\bar{\pi}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{1}$$

where

$$Z = p(\mathbf{y}) = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2}$$

$$= \int_{\Theta} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad \text{with} \quad \pi(\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}), \tag{3}$$

is the so-called marginal likelihood, a.k.a., Bayesian evidence. This quantity is important for model selection purpose, as we show below. However, usually $Z = p(\mathbf{y})$ is unknown and difficult to approximate. We can evaluate the integrand, i.e., unnormalized target (posterior) function, $\pi(\boldsymbol{\theta}$. Clearly, note that $\bar{\pi}(\boldsymbol{\theta}) = \frac{1}{Z}\pi(\boldsymbol{\theta})$. Several methods have been proposed to compute the marginal likelihood Z, and virtually all of them rely on importance sampling (IS) identities [6, 5, 12, 21]. In contrast, the nested sampling (NS) algorithm is primarily based on a different class of representations, often referred to as vertical likelihood representations [28]. The connections and similarities between NS and related IS schemes are discussed in Section 6.

3 One-dimensional representations of the marginal likelihood

In this section, we present an alternative approach based on the Lebesgue representation of the integral defining the marginal likelihood Z. We begin by deriving two equivalent one-dimensional integral formulations of Z, and then discuss how these representations can be exploited through the application of one-dimensional quadrature methods. The practical implementation of such quadrature rules, however, is not straightforward. An appropriate and carefully design of the nested sampling (NS) scheme will be therefore required to make their use effective.

3.1 First one-dimensional representation

The D_θ-dimensional integral $Z = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$ can be turned into a one-dimensional integral using an extended space representation. Namely, we can write

$$
\begin{aligned}
Z &= \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}, && (4)\\
&= \int_\Theta g(\boldsymbol{\theta}) d\boldsymbol{\theta} \int_0^{\ell(\mathbf{y}|\boldsymbol{\theta})} d\lambda, \quad \text{(extended space representation)} && (5)\\
&= \int_\Theta g(\boldsymbol{\theta}) d\boldsymbol{\theta} \int_0^{\infty} \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\} d\lambda, && (6)
\end{aligned}
$$

where $\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}$ is an indicator function which is 1 if $\lambda \in [0, \ell(\mathbf{y}|\boldsymbol{\theta})]$ and 0 otherwise. Switching the integration order, we obtain

$$
\boxed{
\begin{aligned}
Z &= \int_0^{\infty} d\lambda \int_\Theta g(\boldsymbol{\theta}) \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\} d\boldsymbol{\theta}\\
&= \int_0^{\infty} d\lambda \int_{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda} g(\boldsymbol{\theta}) d\boldsymbol{\theta}\\
&= \int_0^{\infty} Z(\lambda) d\lambda,\\
&= \int_0^{\sup \ell(\mathbf{y}|\boldsymbol{\theta})} Z(\lambda) d\lambda,
\end{aligned}
} \qquad (7)
$$

where we have set

$$
\boxed{Z(\lambda) = \int_{\lambda < \ell(\mathbf{y}|\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{P}\left(\lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\right), \quad \text{where} \quad \boldsymbol{\theta} \sim g(\boldsymbol{\theta}).} \qquad (8)
$$

Remark 1. *$Z(\lambda)$ represents a normalized area, with $Z(\lambda) \in [0, 1]$ (as shown in Figure 2). Thus, we have $Z(0) = 1$ and $Z(\lambda') = 0$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$, and is also a non-increasing function, namely, is a survival function.*

In Eq. (7), we have also assumed that $\ell(\mathbf{y}|\boldsymbol{\theta})$ is bounded so the limit of integration is $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$.

Below, we define several variables and sampling procedures required for the proper understanding of the nested sampling algorithm.

3.2 Second one-dimensional representation

Now let consider a specific *area* value $a = Z(\lambda)$. The inverse function

$$\boxed{\Lambda(a) = Z^{-1}(a) = \sup\{\lambda : Z(\lambda) > a\},} \tag{9}$$

is also non-increasing. Note that $Z(\lambda) > a$ if and only if $\lambda < \Lambda(a)$. We have also $\Lambda(0) = \sup \ell(\mathbf{y}|\boldsymbol{\theta})$ and $\Lambda(1) = 0$, i.e., $\Lambda(a) : [0, 1] \to [0, \sup \ell(\mathbf{y}|\boldsymbol{\theta}))$. Then, we can write

$$\begin{aligned}
Z &= \int_0^\infty Z(\lambda) d\lambda \\
&= \int_0^\infty d\lambda \int_0^1 \mathbb{I}\{a < Z(\lambda)\} da \qquad \text{(again the extended space “trick”)} \\
&= \int_0^1 da \int_0^\infty \mathbb{I}\{u < Z(\lambda)\} d\lambda \qquad \text{(switching the integration order)} \\
&= \int_0^1 da \int_0^\infty \mathbb{I}\{\lambda < \Lambda(a)\} d\lambda \qquad \text{(using } Z(\lambda) > a \iff \lambda < \Lambda(a)\text{)} \\
&= \int_0^1 \Lambda(a) da.
\end{aligned} \tag{10}$$

Remark 2. *Hence, we have obtained two of the marginal likelihood Z in Eqs.* (7)-(10)*, that is generally defined by an highly-multidimensional integral given in Eq.* (4)*.*

Note that functions $Z(\lambda)$ and $\Lambda(a)$ take into account both the prior density and likelihood function. See Figure 1 for a graphical representation of these two one-dimensional functions.

3.3 Summary and possible quadratures

Previously, we have obtained two one-dimensional integrals for expressing the Bayesian evidence Z,

$$\boxed{Z = \int_0^{\sup \ell(\mathbf{y}|\boldsymbol{\theta})} Z(\lambda) d\lambda = \int_0^1 \Lambda(a) da.} \tag{11}$$

The two integrand functions $a = Z(\lambda)$ and $\lambda = \Lambda(a)$ are graphically represented in Figure 1. We have expressed the quantity Z as an integral of a one-dimensional function in bounded domain, we could think of applying simple quadrature: choose a grid of points in $[0, \sup \ell(\mathbf{y}|\boldsymbol{\theta})]$ ($\lambda_i > \lambda_{i-1}$) or in $[0, 1]$ ($a_i > a_{i-1}$), evaluate $Z(\lambda)$ or $\Lambda(a)$ and use the quadrature formulas of the form:

$$\widehat{Z} = \sum_{i=1}^{I} (\lambda_i - \lambda_{i-1}) Z(\lambda_i), \text{ or} \tag{12}$$

$$\widehat{Z} = \sum_{i=1}^{I} (a_i - a_{i-1}) \Lambda(a_i). \tag{13}$$

However, these simple approaches are difficult to be applied since (i) the functions $Z(\lambda)$ and $\Lambda(a)$ are intractable in most cases and (ii) they change much more rapidly over their domains than does $\pi(\boldsymbol{\theta}|\mathbf{y}) = \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta})$, hence the quadrature approximation can have very bad performance, unless the grid of points is chosen with extreme care.

In the remainder of this work, we describe in detail the key concepts, variables, procedures, and results required for a proper introduction to the nested sampling method.

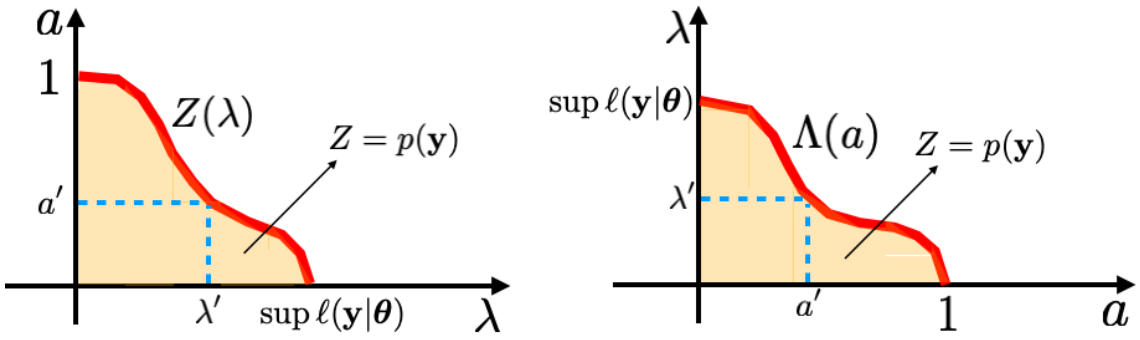


Figure 1: Graphical examples of the two one-dimensional functions $Z(\lambda)$ and $\Lambda(a)$. The area below the curve is in both cases the marginal likelihood Z. The marginal likelihood is generally expressed by a multi-dimensional integral, i.e., $Z = \int_\Theta \ell(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

4 Additional theoretical foundations of NS

4.1 On the survival function $Z(\lambda)$ and related sampling procedures

The function above $Z(\lambda) : \mathbb{R}^+ \to [0, 1]$ is the mass of the prior restricted to the set $\{\boldsymbol{\theta} : \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}$.

Note also that

$$\boxed{Z(\lambda) = \mathbb{P}\left(\lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\right), \quad \text{where} \quad \boldsymbol{\theta} \sim g(\boldsymbol{\theta}).} \tag{14}$$

Moreover, we have that $Z(\lambda) \in [0, 1]$ with $Z(0) = 1$ and $Z(\lambda') = 0$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$, and it is also a non-increasing function. Therefore, $Z(\lambda)$ is a *survival function*, while

$$\boxed{F(\lambda) = 1 - Z(\lambda) = \mathbb{P}\left(\lambda > \ell(\mathbf{y}|\boldsymbol{\theta})\right) = \mathbb{P}\left(L < \lambda\right),} \tag{15}$$

is the cumulative function of a random variable $L = \ell(\mathbf{y}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$ [24, 29]. Note that $F(0) = 0$ and $F(\lambda') = 1$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$.

4.2 Sampling according to $Z(\lambda)$

The following procedure generates samples λ_n from $F(\lambda)$ in Eq. (15):

1. Draw $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta})$, for $n = 1, ..., N$.

2. Set $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$, for all $n = 1, ..., N$.

Recalling the inversion method [24, Chapter 2], note also that the corresponding values

$$b_n = F(\lambda_n) \sim \mathcal{U}([0,1]), \tag{16}$$

i.e., they are uniformly distributed in $[0,1]$. Recall that, if $U \sim \mathcal{U}([0,1])$, $V = 1 - U$ is also uniformly distributed $\mathcal{U}([0,1])$, then we also have

$$a_n = Z(\lambda_n) = 1 - F(\lambda_n) \sim \mathcal{U}([0,1]). \tag{17}$$

In summary, finally we have that

$$\boxed{\begin{array}{rl} \text{if} & \boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}), \\ \text{and} & \lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \sim F(\lambda), \\ \text{then} & a_n = Z(\lambda_n) \sim \mathcal{U}([0,1]). \end{array}} \tag{18}$$

4.3 The likelihood-truncated prior density $g(\cdot|\lambda)$

Note that $Z(\lambda)$ represents also the normalizing constant of the following truncated prior pdf, i.e.,

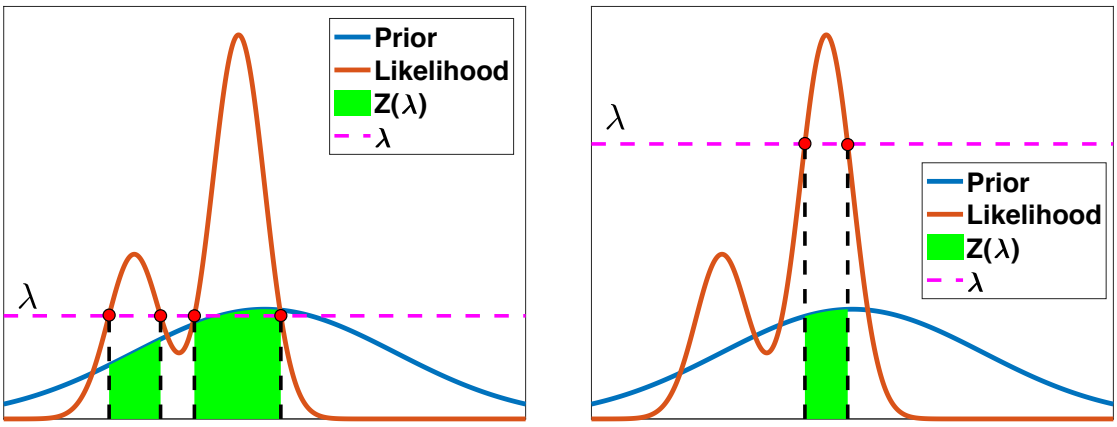
$$\boxed{g(\boldsymbol{\theta}|\lambda) = \frac{1}{Z(\lambda)} \mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\} g(\boldsymbol{\theta}) = \begin{cases} g(\boldsymbol{\theta}), & \text{if } \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda, \\ 0, & \text{otherwise,} \end{cases}} \tag{19}$$

Clearly, we have the two extreme cases:

$$g(\boldsymbol{\theta}|0) = g(\boldsymbol{\theta}), \tag{20}$$

$$g(\boldsymbol{\theta}|\lambda_{\texttt{max}}) = \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{ML}) \quad \text{for} \quad \lambda_{\texttt{max}} = \ell(\mathbf{y}|\widehat{\boldsymbol{\theta}}_{ML}), \tag{21}$$

where $\widehat{\boldsymbol{\theta}}_{ML} = \arg\max \ell(\mathbf{y}|\boldsymbol{\theta})$, i.e., is the maximum likelihood estimator. Two graphical examples of $g(\boldsymbol{\theta}|\lambda)$ and $Z(\lambda)$ are given in Figure 2. The red line depicts the likelihood function, and the blue line shows the complete prior density $g(\boldsymbol{\theta})$. The portions of the blue line that define the green areas represent the truncated prior $g(\boldsymbol{\theta}|\lambda)$.



(a) (b)

Figure 2: Two examples of the area below the truncated prior $g(\boldsymbol{\theta}|\lambda)$, that is represented by the function $Z(\lambda)$ (given by the green areas). The red line depicts the likelihood function, and the blue line shows the complete prior density $g(\boldsymbol{\theta})$. The portions of the blue line that define the green areas represent the truncated prior $g(\boldsymbol{\theta}|\lambda)$. Note that in figure (b) the value of λ is greater than in figure (a), so that the area $Z(\lambda)$ decreases with respect to Figure (a). Clearly, if we assume a λ value bigger than the maximum of the likelihood, then we would have $Z(\lambda) = 0$.

Remark 3. *It is important to emphasize that the truncation of the prior density is defined through the likelihood function, as shown in Figure 2. In particular, determining the support of the truncated prior - that is, the region where $g(\boldsymbol{\theta}|\lambda) > 0$ is well defined and strictly positive - would in general require inverting the likelihood function (which is generally impossible or computational complex).*

4.4 Sampling from the truncated prior $g(\cdot|\lambda)$

Given a fixed value $\lambda_0 \geq 0$, in order to generate samples from $g(\boldsymbol{\theta}|\lambda_0)$ one alternative is to use an MCMC procedure. However, as an example, the following acceptance-rejection procedure can be employed [24]:

1. For $n = 1, ..., N$:
   (a) Draw $\boldsymbol{\theta}' \sim g(\boldsymbol{\theta})$.
   (b) if $\ell(\mathbf{y}|\boldsymbol{\theta}') > \lambda_0$ then set $\boldsymbol{\theta}_{n|0} = \boldsymbol{\theta}'$ and $\lambda_{n|0} = \ell(\mathbf{y}|\boldsymbol{\theta}')$.
   (c) if $\ell(\mathbf{y}|\boldsymbol{\theta}') \leq \lambda_0$, then reject $\boldsymbol{\theta}'$ and repeat from step 1(a).
2. Return $\{\boldsymbol{\theta}_{n|0}\}_{n=1}^N$ and $\{\lambda_{n|0}\}_{n=1}^N$, where all $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_0)$ and all $\lambda_{n|0} \sim F(\lambda|\lambda_0)$, given below.

Observe that $\boldsymbol{\theta}_{n|0} \sim g(\boldsymbol{\theta}|\lambda_0)$, for all $n = 1, ..., N$, and the probability of accepting a generated sample $\boldsymbol{\theta}'$ is exactly $Z(\lambda)$. The values $\lambda_{n|0} = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$ where $\boldsymbol{\theta}_{n|0} \sim g(\boldsymbol{\theta}|\lambda_0)$, have the following *truncated* cumulative function

$$F(\lambda|\lambda_0) = \frac{F(\lambda) - F(\lambda_0)}{1 - F(\lambda_0)} = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda_0), \quad \text{with } \lambda \geq \lambda_0, \text{ and } \boldsymbol{\theta} \sim g(\boldsymbol{\theta}|\lambda_0), \tag{22}$$

i.e., we can write $\lambda_n \sim F(\lambda|\lambda_0)$. Namely, considering the steps above, (a) draw $\boldsymbol{\theta}' \sim g(\boldsymbol{\theta})$ and (b) if $\ell(\mathbf{y}|\boldsymbol{\theta}') > \lambda_0$ then set $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}')$, we can generate samples according to the truncated cumulative function $F(\lambda|\lambda_0)$. Setting $\lambda_0 = 0$, we have $F(\lambda|0) = F(\lambda)$ that is the cumulative function defined in Eq. (15). Recall that $F(0) = 0$ and $F(\lambda') = 1$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$. Similarly, we have $F(\lambda'|\lambda_0) = 0$ for all $\lambda' < \lambda_0$ and $F(\lambda'|\lambda_0) = 1$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$.

4.5 Distribution of $a_{n|0} = Z(\lambda_{n|0})$ with $\lambda_{n|0} \sim F(\lambda|\lambda_0)$

Recall that $Z(\lambda)$ is non-increasing, then $Z(\lambda) \leq Z(\lambda_0)$ if $\lambda_0 \leq \lambda$. Moreover, recall that with $Z(0) = 1$ and $Z(\lambda') = 0$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$. With similar arguments of Section 4.2 and

similarly to Eq. (18), we can write

$$\boxed{\begin{array}{rll} \text{if} & \boldsymbol{\theta}_{n|0} \sim g(\boldsymbol{\theta}|\lambda_0), & \\ \text{and} & \lambda_{n|0} = \ell(\mathbf{y}|\boldsymbol{\theta}_{n|0}) \sim F(\lambda|\lambda_0), & \\ \text{then} & a_{n|0} = Z(\lambda_{n|0}) \sim \mathcal{U}([0, a_0]), & \text{with } a_0 = Z(\lambda_0) \leq 1. \end{array}} \tag{23}$$

Clearly, the expression above ensured that $a_{n|0} = Z(\lambda_n) \leq a_0 = Z(\lambda_0)$, i.e., $a_{n|0} \leq a_0$ (as expected since $Z(\lambda)$ is non-increasing and $\lambda_n \geq \lambda_0$). Since $a_{n|0} \sim \mathcal{U}([0, a_0])$, we can also write:

$$\boxed{\widetilde{a}_n = \frac{a_{n|0}}{a_0} \sim \mathcal{U}([0,1]), \quad \forall n = 1, ..., N.} \tag{24}$$

4.6 Distribution of $\widetilde{a}_{\texttt{max}}$

Let us consider $\lambda_{1|0}, ..., \lambda_{N|0} \sim F(\lambda|\lambda_0)$ and the minimum and maximum values

$$\lambda_{\texttt{min}} = \min_n \lambda_{n|0}, \quad a_{\texttt{max}|0} = Z(\lambda_{\texttt{min}}), \quad \text{and} \quad \widetilde{a}_{\texttt{max}} = \frac{a_{\texttt{max}|0}}{a_0} = \frac{Z(\lambda_{\texttt{min}})}{Z(\lambda_0)}. \tag{25}$$

Let us recall $\widetilde{a}_n = \frac{a_{n|0}}{a_0} \sim \mathcal{U}([0,1])$. Then, note that $\widetilde{a}_{\texttt{max}}$ is maximum of N uniform random variables

$$\widetilde{a}_1, ..., \widetilde{a}_N \sim \mathcal{U}([0,1]).$$

Then it is well-known that the cumulative distribution of the maximum value

$$\widetilde{a}_{\texttt{max}} = \max_n \widetilde{a}_n \sim \mathcal{B}(N, 1),$$

is distributed according to a Beta distribution $\mathcal{B}(N,1)$, i.e., with cumulative function $F_{\texttt{max}}(\widetilde{a}) = \widetilde{a}^N$ and density $f_{\texttt{max}}(\widetilde{a}) = \frac{dF_{\texttt{max}}(\widetilde{a})}{d\widetilde{a}} = N\widetilde{a}^{N-1}$ [24, Section 2.3.6]. In summary, we have

$$\boxed{\begin{array}{rl} & \widetilde{a}_{\texttt{max}} = \dfrac{Z(\lambda_{\texttt{min}})}{Z(\lambda_0)} \sim \mathcal{B}(N,1), \\ \text{where} & \lambda_{\texttt{min}} = \min\limits_n \lambda_{n|0}, \\ \text{and} & \lambda_{1|0}, ..., \lambda_{N|0} \sim F(\lambda|\lambda_0). \end{array}} \tag{26}$$

This result is important for deriving the standard version of the nested sampling method, described in the next section. A summary of the relationships presented above is provided in Table 1.

Table 1: Summary of the relationships among some random variables introduced above.

Sections	**Relationships**
4.5	If $\boldsymbol{\theta}_{n\|0} \sim g(\boldsymbol{\theta}\|\lambda_0)$, and $\lambda_{n\|0} = \ell(\mathbf{y}\|\boldsymbol{\theta}_{n\|0}) \sim F(\lambda\|\lambda_0)$, $\widetilde{a}_n = \dfrac{a_{n\|0}}{a_0} = \dfrac{Z(\lambda_{n\|0})}{Z(\lambda_0)} \sim \mathcal{U}([0,1])$.
4.6	If $\boldsymbol{\theta}_{n\|0} \sim g(\boldsymbol{\theta}\|\lambda_0)$, and $\lambda_{n\|0} = \ell(\mathbf{y}\|\boldsymbol{\theta}_{n\|0}) \sim F(\lambda\|\lambda_0)$, $n = 1, ..., N$, $\lambda_{\texttt{min}} = \min_n \lambda_{n\|0}$, $\widetilde{a}_{\texttt{max}} = \max \widetilde{a}_n = \dfrac{Z(\lambda_{\texttt{min}})}{Z(\lambda_0)} \sim \mathcal{B}(N,1)$.

5 A detailed description of nested sampling (NS)

5.1 Form of the NS estimator

Nested sampling (NS) is a technique for estimating the marginal likelihood that exploits the second identity in (11) [31, 9, 28]. Nested Sampling estimates Z by a quadrature using nodes (in *decreasing* order),

$$0 < a_{\texttt{max}}^{(I)} < \cdots < a_{\texttt{max}}^{(1)} < a_{\texttt{max}}^{(0)} = 1$$

and the quadrature formula

$$\boxed{\widehat{Z} = \sum_{i=1}^{I} (a_{\texttt{max}}^{(i-1)} - a_{\texttt{max}}^{(i)}) \Lambda(a_{\texttt{max}}^{(i)}) = \sum_{i=1}^{I} (a_{\texttt{max}}^{(i-1)} - a_{\texttt{max}}^{(i)}) \lambda_{\texttt{min}}^{(i)},} \tag{27}$$

with $a_{\texttt{max}}^{(0)} = 1$. Furthermore, we will see that the NS construction yields that

$$\boxed{\lim_{I \to \infty} a_{\texttt{max}}^{(I)} = a_{\texttt{max}}^{(\infty)} = 0.} \tag{28}$$

Remark 4. *Recall that $\lambda = \ell(\mathbf{y}|\boldsymbol{\theta})$'s represent likelihood values, and a's represent normalized area values contained in $[0, 1]$.*

In Eq. (27), we have to specify the grid points $a_{\texttt{max}}^{(i)}$'s (possibly well-located, with a suitable strategy) and the corresponding values $\lambda_{\texttt{min}}^{(i)} = \Lambda(a_{\texttt{max}}^{(i)})$. Recall that the function $\Lambda(a)$, and its inverse $a = \Lambda^{-1}(\lambda) = Z(\lambda)$, are generally intractable, so that it is not even possible to evaluate $\Lambda(a)$ at a grid of chosen $a_{\texttt{max}}^{(i)}$'s.

Remark 5. *The nested sampling algorithm works in the other way around: it suitably selects the ordinates $\lambda_{min}^{(i)}$'s (likelihood values) and find some approximations $\widehat{a}_i$'s of the corresponding values $a_{max}^{(i)} = Z(\lambda_{min}^{(i)})$. This is possible since the distribution of $a_{max}^{(i)}$ is known (see Section 4.6).*

5.2 Choice of $\lambda_{\mathtt{min}}^{(i)}$ and $a_{\mathtt{max}}^{(i)}$ in nested sampling

In this section, we adopt a slightly simplified notation compared to the previous part of the work, as nested sampling involves additional indices that may otherwise obscure the exposition and hinder the reader's understanding. Nested sampling employs an iterative procedure in order to generate an *increasing* sequence of likelihood ordinates $\lambda_{\mathtt{min}}^{(i)}$, $i = 1, ..., I$, such that

$$\lambda_{\mathtt{min}}^{(1)} < \lambda_{\mathtt{min}}^{(2)} < \lambda_{\mathtt{min}}^{(3)} .... < \lambda_{\mathtt{min}}^{(I)}. \tag{29}$$

Table 2 provides a compact and complete description of the standard NS algorithm that is based on the sampling of the truncated prior pdf $g(\boldsymbol{\theta}|\lambda_{\mathtt{min}}^{(i-1)})$ (see Section 4.3 for the related details), where i denotes an iteration index. The nested sampling approach is explained with more details below:

- At the first iteration ($i = 1$), we set $\lambda_{\mathtt{min}}^{(0)} = 0$ and $a_{\mathtt{max}}^{(0)} = Z(\lambda_{\mathtt{min}}^{(0)}) = 1$. Then, N samples are drawn from the prior $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_{\mathtt{min}}^{(0)}) = g(\boldsymbol{\theta})$ obtaining an (initial) cloud

$$\mathcal{P} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N\} = \{\boldsymbol{\theta}_n\}_{n=1}^N, \tag{30}$$

  often called *set of "live points"* and then set $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$, i.e., $\{\lambda_n\}_{n=1}^N \sim F(\lambda)$ as shown in Section 4.1. Thus, the first ordinate is chosen as

$$\lambda_{\mathtt{min}}^{(1)} = \min_n \lambda_n = \min_n \ell(\mathbf{y}|\boldsymbol{\theta}_n) = \min_{\boldsymbol{\theta}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P}) = \ell(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{rem}}^{(1)}),$$

  where $\boldsymbol{\theta}_{\mathtt{rem}}^{(1)} = \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$ is also removed from $\mathcal{P}$, i.e., $\mathcal{P} = \mathcal{P}\backslash\{\boldsymbol{\theta}_{\mathtt{rem}}^{(1)}\}$ (now $|\mathcal{P}| = N - 1$). Moreover, since $\{\lambda_n\}_{n=1}^N \sim F(\lambda)$, using the result in Eq. (26), we have that

$$\widetilde{a}_{\mathtt{max}}^{(1)} = \frac{a_{\mathtt{max}}^{(1)}}{a_{\mathtt{max}}^{(0)}} = \frac{Z(\lambda_{\mathtt{min}}^{(1)})}{Z(\lambda_{\mathtt{min}}^{(0)})} \sim \mathcal{B}(N, 1).$$

Since $a_{\texttt{max}}^{(0)} = Z(\lambda_{\texttt{min}}^{(0)}) = 1$, then $\widetilde{a}_{\texttt{max}}^{(1)} = a_{\texttt{max}}^{(1)} \sim \mathcal{B}(N, 1)$.

- At a generic i-th iteration ($i \geq 2$), a unique additional sample $\boldsymbol{\theta}'$ is drawn from the truncated prior $g(\boldsymbol{\theta}|\lambda_{\texttt{min}}^{(i-1)})$ and added to the current cloud of samples, i.e., $\mathcal{P} = \mathcal{P} \cup \boldsymbol{\theta}'$ (now again $|\mathcal{P}| = N$). First of all, note that the value $\lambda' = \lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}')$ is distributed as $F(\lambda|\lambda_{\texttt{min}}^{(i-1)})$ (see Section 4.3). More precisely, note that all the N ordinate (likelihood) values

$$\{\lambda_n\}_{n=1}^N = \ell(\mathbf{y}|\mathcal{P}) = \{\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \;\text{ for all }\; \boldsymbol{\theta}_n \in \mathcal{P}\}$$

are distributed as $F(\lambda|\lambda_{\texttt{min}}^{(i-1)})$, i.e., $\{\lambda_n\}_{n=1}^N \sim F(\lambda|\lambda_{\texttt{min}}^{(i-1)})$. This is due to how the population $\mathcal{P}$ has been built in the previous iterations. Then, we choose the new minimum value as

$$\lambda_{\texttt{min}}^{(i)} = \min_n \lambda_n = \min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\mathcal{P}) = \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)}).$$

We remove again the corresponding sample $\boldsymbol{\theta}_{\texttt{rem}}^{(i)} = \arg\min_{\boldsymbol{\theta} \in \mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$, i.e., we set $\mathcal{P} = \mathcal{P} \backslash \{\boldsymbol{\theta}_{\texttt{rem}}^{(i)}\}$ and the procedure is repeated. Moreover, since $\lambda_{\texttt{min}}^{(i)}$ is the minimum value of $\{\lambda_1, ..., \lambda_N\} \sim F(\lambda|\lambda_{\texttt{min}}^{(i-1)})$, in Section 4.6 we have seen that

$$\widetilde{a}_{\texttt{max}}^{(i)} = \frac{a_{\texttt{max}}^{(i)}}{a_{\texttt{max}}^{(i-1)}} = \frac{Z(\lambda_{\texttt{min}}^{(i)})}{Z(\lambda_{\texttt{min}}^{(i-1)})} \sim \mathcal{B}(N, 1), \tag{31}$$

where we have used Eq. (26). Note that we have also found the recursion among the following random variables,

$$a_{\texttt{max}}^{(i)} = \widetilde{a}_{\texttt{max}}^{(i)} a_{\texttt{max}}^{(i-1)}, \tag{32}$$

for $i = 1, ..., I$ and $a_{\texttt{max}}^{(0)} = 1$.

- The random value $\widetilde{a}_{\texttt{max}}^{(i)}$ could be estimated and replaced with the expected value of the Beta

distribution $\mathcal{B}(N,1)$, i.e.,

$$\widetilde{a}_{\mathtt{max}}^{(i)} \approx \mathbb{E}[\mathcal{B}(N,1)] = \frac{N}{N+1} \approx \exp\left(-\frac{1}{N}\right). \tag{33}$$

where $\mathbb{E}[\mathcal{B}(N,1)] = \frac{N}{N+1}$ is used as estimator, and $\exp\left(-\frac{1}{N}\right)$ becomes a very good approximation of $\mathbb{E}[\mathcal{B}(N,1)]$ as N grows. In that case, *assuming additionally the independence of the live points in the population* (i.e., we are using the property of the expectation valid for independent variables, $\mathbb{E}[a_{\mathtt{max}}^{(i)}] = \mathbb{E}[\widetilde{a}_{\mathtt{max}}^{(i)}]\mathbb{E}[a_{\mathtt{max}}^{(i-1)}]$), the recursion above becomes

$$a_{\mathtt{max}}^{(i)} \approx \exp\left(-\frac{1}{N}\right) a_{\mathtt{max}}^{(i-1)} = \exp\left(-\frac{i}{N}\right). \tag{34}$$

Then, denoting $\widehat{a}_i = \exp\left(-\frac{i}{N}\right)$, we can use $\widehat{a}_i$ as an approximation of $a_{\mathtt{max}}^{(i)}$.

The algorithm is given in Table 2. Figure 3 illustrates, with $N = 3$, the removal and replacement steps of the live points within the population $\mathcal{P}$ between the two consecutive iterations. Note that the NS procedure tends to dynamically add live points near to the main mode or, more generally, to areas of high probability mass. To properly understand the definition and sampling of the truncated prior density, we recommend accurately examining Figure 2.

Remark 6. *The intuition behind the iterative approach above is to accumulate more ordinates* λ_i *close to the* $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$. *They are also more dense around* $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$. *Moreover, using this scheme, we can employ* $\widehat{a}_i = \exp\left(-\frac{i}{N}\right)$ *as an approximation of* $a_{max}^{(i)}$.

Remark 7. *An implicit optimization of the likelihood function is performed in the nested sampling algorithm. All the value* $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n)$ *with* $\boldsymbol{\theta}_n \in \mathcal{P}$ *approaches the value* $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$.

Remark 8. *Note that with respect to the* θ*-space, the NS method is a population sampler where a cloud of points* $\mathcal{P} = \{\boldsymbol{\theta}_n\}_{n=1}^N$ *that changes with the iterations, adding and removing one point per iteration.*

Table 2: The standard nested sampling procedure.

1. Choose the number of points per iterations N, the total number of iterations I (or an alternative stopping rule), and set $\widehat{a}_0 = 1$.

2. Draw $\{\boldsymbol{\theta}_n\}_{n=1}^N \sim g(\boldsymbol{\theta})$ and define the set $\mathcal{P} = \{\boldsymbol{\theta}_n\}_{n=1}^N$. Let us also define the notation

$$\ell(\mathbf{y}|\mathcal{P}) = \{\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \text{ for all } \boldsymbol{\theta}_n \in \mathcal{P}\}, \tag{35}$$

3. Set $\lambda_{\texttt{min}}^{(1)} = \min_{\boldsymbol{\theta}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P}) = \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(1)})$ where $\boldsymbol{\theta}_{\texttt{rem}}^{(1)} = \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$.

4. Set $\mathcal{P} = \mathcal{P}\backslash\{\boldsymbol{\theta}_{\texttt{rem}}^{(1)}\}$, i.e., remove $\boldsymbol{\theta}_{\texttt{rem}}^{(1)}$ from $\mathcal{P}$.

5. Find an approximation $\widehat{a}_1$ of $a_{\texttt{max}}^{(1)} = Z(\lambda_{\texttt{min}}^{(1)})$. One usual choice is $\widehat{a}_1 = \exp\left(-\frac{1}{N}\right)$.

6. For $i = 2, .., I$:

   (a) Draw $\boldsymbol{\theta}' \sim g(\boldsymbol{\theta}|\lambda_{\texttt{min}}^{(i-1)})$ and add to the current cloud of samples, i.e., $\mathcal{P} = \mathcal{P} \cup \boldsymbol{\theta}'$.

   (b) Set $\lambda_{\texttt{min}}^{(i)} = \min_{\boldsymbol{\theta}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P}) = \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)})$ where $\boldsymbol{\theta}_{\texttt{rem}}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$.

   (c) Set $\mathcal{P} = \mathcal{P}\backslash\{\boldsymbol{\theta}_{\texttt{rem}}^{(i)}\}$, i.e., remove $\boldsymbol{\theta}_{\texttt{rem}}^{(i)}$ from the set of *live points* $\mathcal{P}$.

   (d) Find an approximation $\widehat{a}_i$ of $a_{\texttt{max}}^{(i)} = Z(\lambda_{\texttt{min}}^{(i)})$. One usual choice is

$$\widehat{a}_i = \exp\left(-\frac{i}{N}\right) \approx a_{\texttt{max}}^{(i)}, \tag{36}$$

   The rationale behind this choice is explained in the section above.

7. Return

$$\widehat{Z} = \sum_{i=1}^{I} (\widehat{a}_{i-1} - \widehat{a}_i)\lambda_{\texttt{min}}^{(i)} = \sum_{i=1}^{I} (e^{-\frac{i-1}{N}} - e^{-\frac{i}{N}})\lambda_{\texttt{min}}^{(i)}. \tag{37}$$

6 Relationship with importance sampling (IS)

We can rewrite Eq. (27) as

$$\boxed{\begin{aligned} \widehat{Z} &= \sum_{i=1}^{I} \underbrace{(a_{\texttt{max}}^{(i-1)} - a_{\texttt{max}}^{(i)})}_{\gamma_i} \underbrace{\lambda_{\texttt{min}}^{(i)}}_{\ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)})}, \\ \widehat{Z} &= \sum_{i=1}^{I} \gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)}), \end{aligned}} \tag{38}$$

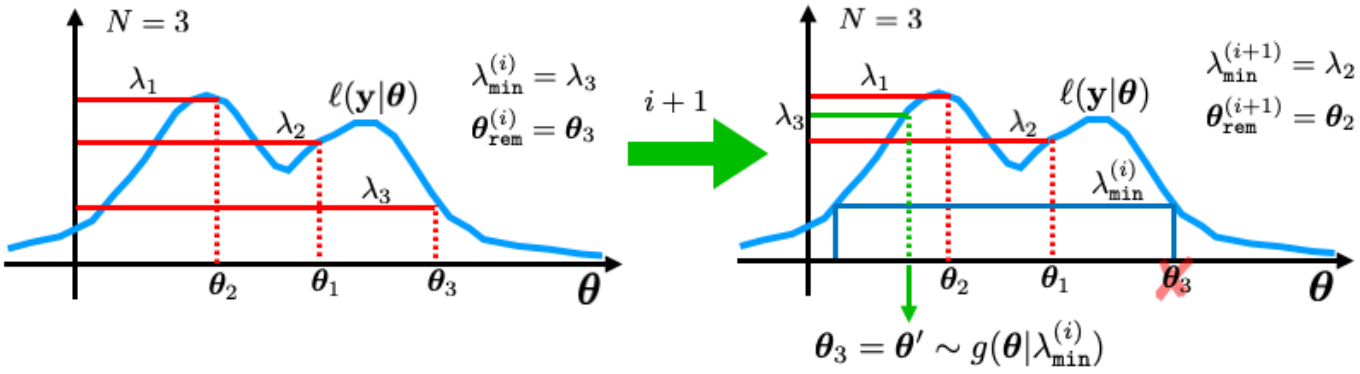


Figure 3: Graphical representation of the NS procedure from the i-th iteration to the $(i+1)$-th iteration, with $N = 3$ live points within the population $\mathcal{P} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$. The blue line represents the likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta})$. At i-th iteration, the point $\boldsymbol{\theta}_3$ has the lowest likelihood value, i.e., $\lambda^{(i)}_{\texttt{min}} = \lambda_3 = \ell(\mathbf{y}|\boldsymbol{\theta}_3)$. Hence, $\boldsymbol{\theta}_3$ is removed from the population $\mathcal{P}$ and, at the next $i+1$-th iteration, a new sample is generated $\boldsymbol{\theta}_3 = \boldsymbol{\theta}' \sim g(\boldsymbol{\theta}|\lambda^{(i-1)}_{\texttt{min}})$ and add to the set $\mathcal{P}$ in order to keep fixed the number of live points $N = |\mathcal{P}| = 3$. The new sample generation, at the $i+1$-th iteration, is done according to the truncated prior $g(\boldsymbol{\theta}|\lambda^{(i-1)}_{\texttt{min}})$ (see Figure 2). New live points tend to be added in regions of high probability mass (near to the global mode).

where we have set $\gamma_i = a^{(i-1)}_{\texttt{max}} - a^{(i)}_{\texttt{max}} > 0$ and $\gamma_i \in [0, 1]$ for all i. Indeed, recall that $a^{(i)}_{\texttt{max}}$ are positive and decreasing values and $a^{(i)}_{\texttt{max}} \in [0, 1]$ (they represent normalized areas). Furthermore, their sum is approximately 1, i.e.,

$$\begin{aligned}\sum_{i=1}^{I} \gamma_i &= (a^{(0)}_{\texttt{max}} - a^{(1)}_{\texttt{max}}) + (a^{(1)}_{\texttt{max}} - a^{(2)}_{\texttt{max}}) + (a^{(2)}_{\texttt{max}} - a^{(3)}_{\texttt{max}}) + ... + (a^{(I-1)}_{\texttt{max}} - a^{(I)}_{\texttt{max}}), \\ &= a^{(0)}_{\texttt{max}} - a^{(I)}_{\texttt{max}}, \\ &= 1 - a^{(I)}_{\texttt{max}}. \end{aligned} \tag{39}$$

if $I \to \infty$, we have $\sum_{i=1}^{I} \gamma_i \approx 1$. where we have used $\lim_{I\to\infty} a^{(I)}_{\texttt{max}} = a^{(\infty)}_{\texttt{max}} = 0$. From the expression above, it is apparent that $\widehat{Z}$ is a linear (asymptotically convex, for $I \to \infty$) combination of weights γ_i. In [21], the authors describe a similar importance sampling (IS) estimator for the marginal

likelihood Z:

$$\boxed{\widehat{Z} = \sum_{i=1}^{I} \rho_i \ell(\mathbf{y}|\boldsymbol{\theta}_i), \qquad \{\boldsymbol{\theta}_i\}_{i=1}^{I} \sim q(\boldsymbol{\theta}),} \tag{40}$$

where $q(\boldsymbol{\theta})$ is a generic proposal density, $\rho_i = \frac{g(\boldsymbol{\theta}_i)}{Iq(\boldsymbol{\theta}_i)}$.

Remark 9. *Hence, the NS estimator can be interpreted as a IS estimator of the form* (40)*, using a "sophisticated" proposal density such that* $\rho_i = \gamma_i = a_{max}^{(i-1)} - a_{max}^{(i)}$*. Clearly, the analytical form of this proposal is not available and we cannot evaluate it. However, we can draw from it using the NS procedure, and the weights are compute by the formula* $\gamma_i = a_{max}^{(i-1)} - a_{max}^{(i)}$*. See also [28] for related comments.*

Remark 10. *The NS weights* $\gamma_i = a_{max}^{(i-1)} - a_{max}^{(i)}$ *represent a partition of prior mass.*

Hence, we can assert that the product $\gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)})$ represents a portion of area below the unnormalized posterior $\pi(\boldsymbol{\theta})$. Indeed, the sum of this products provides an approximation of $Z = \int_\Theta \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$, i.e., $\widehat{Z} = \sum_{i=1}^{I} \gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)})$. Thus, the normalized weights

$$\boxed{\bar{w}_i = \frac{\gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)})}{\widehat{Z}} = \frac{\gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(i)})}{\sum_{k=1}^{I} \gamma_k \ell(\mathbf{y}|\boldsymbol{\theta}_{\texttt{rem}}^{(k)})},} \tag{41}$$

allow the approximation of the posterior measure as

$$\widehat{\pi}(\boldsymbol{\theta}) = \sum_{i=1}^{I} \bar{w}_i \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_{\texttt{rem}}^{(i)}), \tag{42}$$

hence the NS method can be also applied to approximate generic integral involving the posterior distribution as

$$I = \int_\Theta f(\boldsymbol{\theta}) \bar{\pi}(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \widehat{I}_{\texttt{NS}} = \sum_{i=1}^{I} \bar{w}_i f(\boldsymbol{\theta}_{\texttt{rem}}^{(i)}),$$

where is $f(\boldsymbol{\theta})$ is any integrable function.

7 Practical and theoretical limitations in standard NS

In the following, we describe the more critical points in nested sampling:

- Arguably, the most critical task in the implementation of nested sampling is drawing samples from the truncated prior. For this purpose, one can use a rejection sampling or an MCMC scheme. In the first case, we sample from the prior and then accept only the samples $\boldsymbol{\theta}'$ such that $\ell(\mathbf{y}|\boldsymbol{\theta}') > \lambda$. However, as λ grows, its performance deteriorates since the acceptance probability gets smaller and smaller. The MCMC algorithms could also have poor performance due to the sample correlation, especially when the support of the constrained prior is formed by disjoint regions or distant modes [9]. More generally, there are several possible issues: (a) the constrained region can be extremely thin; (b) in high dimension, most of the prior mass is near the boundary; and, as previously noticed, (c) the region may be disconnected (multimodal posteriors). If constrained sampling is imperfect, we can add a bias to the estimator. In practice, the performance depends heavily on the internal sampling method. See also Section 8 below.

- NS is primarily an evidence estimator. Posterior weights can be often highly skewed and, as a consequence, the effective sample size (ESS) can be small (early points have often very small weights) [10, 23, 22]. So NS posterior sampling may be less efficient than well-tuned Monte Carlo for parameter estimation (such as MCMC or IS schemes).

- Moreover, in the derivation of the NS method we have considered different approximations:

  - The value $\widetilde{a}^{(i)}_{\texttt{max}}$ cannot be computed but estimated with the expected value of the Beta distribution $\mathcal{B}(N,1)$, i.e., $\mathbb{E}[\mathcal{B}(N,1)] = \frac{N}{N+1}$.

  - The expected value $\mathbb{E}[\mathcal{B}(N,1)] = \frac{N}{N+1}$ is further replaced and approximated with an exponential function $\exp\left(-\frac{1}{N}\right)$ in Eq. (33). Hence, Eq. (33) contains two approximations. This step could be avoided, keeping directly $\frac{N}{N+1}$. The simplicity

of the final formula $\exp\left(-\frac{i}{N}\right)$ is perhaps the reason of using the approximation $\frac{N}{N+1} \approx \exp\left(-\frac{1}{N}\right)$.

- A further approximation $\mathbb{E}[a_{\texttt{max}}^{(i)}] \approx \mathbb{E}[\widetilde{a}_{\texttt{max}}^{(i)}]\mathbb{E}[a_{\texttt{max}}^{(i-1)}]$ is also implicitly applied in (34) (due to the assumption that "live" points in the population are independent). Clearly, when MCMC is used for the constrained sampling, this assumption is violated. Since MCMC-based constrained sampling is the dominant practical approach, study the impact of this approximation would deserve more attention.
- Additionally, we recall if an MCMC method is run for sampling from the constrained prior, also the likelihood values λ_i are in some sense approximated due to the possible burn-in period of the chain.

Hence, as a summary, the values of the areas $a_{\texttt{max}}^{(i)}$ and the corresponding values $\lambda_{\texttt{min}}^{(i)}$ in the NS estimator (27) are computed approximately (they are analytically intractable/unknown values). Sampling from the likelihood-constrained priors is the crucial point and the main challenge.

8 Advanced NS schemes in literature

Nested Sampling estimators have been extended to a variety of settings. For example, in likelihood-free scenarios - where only an unbiased estimate of the likelihood is available - adapted versions of the NS algorithm have been developed [25]. The combination with importance sampling has been designed in [9].

Furthermore, several strategies have emerged to address the main challenges of NS, such as high-dimensionality, multimodality, and strong parameter degeneracies. In this section, we review the principal variants and approaches. The *ellipsoidal NS* technique, as included in MultiNest implementation [11], approximates the current set of live points by one or more ellipsoids. New samples are drawn uniformly from the union of these ellipsoids and are accepted if they satisfy

the likelihood constraint condition. By employing multiple ellipsoids, the method can effectively handle separated modes. The slice sampling method and the Hamiltonian Monte Carlo (HMC) algorithm has been proposed to be used within NS [15, 2]. The main idea in [2] is to leverage HMC dynamics to traverse the constrained likelihood region more efficiently. By utilizing gradient information, it can reduce the random-walk behavior common in traditional MCMC approaches and enables efficient exploration over long distances in parameter space. This method seems to work well in high-dimensional smooth posteriors and is particularly effective for Bayesian models with differentiable likelihoods. Its limitations include the requirement of gradients and the complexity of correctly handling the likelihood boundary defined by the constrain. In addition, Hamiltonian NS is less robust for multimodal distributions.

The *diffusive Nested Sampling* (DNS) scheme [3] introduces an alternative exploration mechanism in which particles are allowed to diffuse across likelihood levels rather than progressing monotonically toward increasingly constrained regions. Instead of enforcing a strictly inward shrinkage of prior mass, DNS constructs a sequence of likelihood levels that can be explored reversibly. In this framework, particles are not restricted to the current likelihood constraint but may move upward to more constrained levels or downward to less constrained ones, thereby improving mixing across the hierarchy of constrained priors. The up/down move of DNS is achieved stochastically according to Metropolis-Hastings steps. Importantly, the likelihood levels are not defined by all observed likelihood values. Rather, a relatively small number of levels- typically few of them - is maintained. Within each level, the sampled points are used to estimate the average likelihood over the corresponding prior volume, allowing for a more flexible and globally informed approximation of the evidence. This process is reminiscent of tempering and enhances mixing, especially for strongly multimodal distributions. DNS reduces the risk of missing isolated modes that might be overlooked by traditional NS methods. While it seems particularly effective for posteriors with extreme multimodality, it is more computationally complex and can result in higher variance in evidence estimates compared to conventional approaches.

In [14], the author relaxes the strict likelihood constraint by introducing an auxiliary "demon" variable that temporarily absorbs surplus likelihood (interpreted as energy), thereby permitting controlled fluctuations around the imposed likelihood threshold. In [17], the authors have proposed the *dynamic NS* method, i.e., dynamically varying the number of live points throughout the sampling process. This dynamic allocation focuses computational resources on regions where additional samples most effectively reduce uncertainty in the evidence or posterior estimates. Other recent works combining nested sampling and machine learning can be found in [32, 33].

9 Conclusions

The Nested sampling (NS) technique has attracted considerable attention and has been widely applied in the literature. Its success is undeniable, especially in cosmology and astronomy. In our view, NS inherently performs a kind of likelihood optimization, which underlies much of its remarkable success. We consider this feature particularly important, as identifying regions of high likelihood is a critical aspect for the efficiency of all computational sampling methods. Indeed, the use of the gradient on many sampling methods has the same purpose.

The full derivation is complex and relies on several approximations, as discussed in Section 7. Sampling from the likelihood-constrained priors is a central challenge and is far from straightforward [7, 8, 9], with the quality of this constrained sampling being crucial to the method's performance. In this context, the widespread success of NS in the literature is indeed remarkable. In our view, the comprehensive and detailed description presented in this work can not only enhance understanding and appreciation of nested sampling but also serve as a valuable foundation for the development of future improvements and methodological variants. By clarifying both the theoretical principles and the practical challenges of the algorithm, we hope to provide insights that will guide researchers in designing more efficient implementations, exploring novel extensions, and applying NS to an even broader range of scientific problems. In this sense, this work serves

as a tutorial for newcomers to the field and as a critical review for experienced practitioners.

References

[1] G. Ashton, N. Bernstein, J. Buchner, X. Chen, G. Cs√°nyi, A. Fowlie, F. Feroz, M. Griffiths, W. Handley, M. Habeck, E. Higson, M. Hobson, A. Lasenby, D. Parkinson, L. B. P√°rtay, M. Pitkin, D. Schneider, J. S. Speagle, L. South, J. Veitch, P. Wacker, D. J. Wales, and D. Yallup. Nested sampling for physical scientists. *Nature Reviews Methods Primers*, 2(1):39, May 2022.

[2] M. Betancourt. Nested sampling with constrained Hamiltonian Monte Carlo. *AIP Conference Proceedings*, 1305(1):165–172, 03 2011.

[3] B. J. Brewer, L. B. Pártay, and G. Csányi. Diffusive nested sampling. *Statistics and Computing*, 21(4):649–656, 2011.

[4] J. Buchner. Nested sampling methods. *Statistics Surveys*, 17(none):169 – 215, 2023.

[5] M.-H. Chen. Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica*, 59(1):16–29, 2005.

[6] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[7] N. Chopin and C. P. Robert. Comments on nested sampling by John Skilling. *Bayesian Statistics*, 8:491–524, 2007.

[8] N. Chopin and C. P. Robert. Contemplating evidence: properties, extensions of, and alternatives to nested sampling. *Technical Report 2007-46, CEREMADE, Universit Paris Dauphine*, pages 1–26, 2007.

[9] N. Chopin and C. P. Robert. Properties of nested sampling. *Biometrika*, 97(3):741–755, 2010.

[10] V. Elvira, L. Martino, and C. P. Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2022.

[11] F. Feroz, M. P. Hobson, and M. Bridges. Multinest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, 2009.

[12] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607, 2008.

[13] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[14] M. Habeck. Nested sampling with demons. *AIP Conference Proceedings*, 1641(1):121–129, 01 2015.

[15] W. J. Handley, M. P. Hobson, and A. N. Lasenby. Polychord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4384–4398, 2015.

[16] E. Higson, W. Handley, and A. Hobson, M. land Lasenby. Sampling errors in nested sampling parameter estimation. *Bayesian Analysis*, 13(3):873–896, 2018.

[17] E. Higson, W. Handley, M. Hobson, and A. Lasenby. Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. *Statistics and Computing*, 29(5):891–913, 2019.

[18] C. R. Keeton. On statistical uncertainty in nested sampling. *Monthly Notices of the Royal Astronomical Society*, 414(2):1418–1426, 06 2011.

[19] K. H. Knuth, M. Habeck, N. K. Malakar, A. M. Mubeen, and B. Placek. Bayesian evidence and model selection. *Digital Signal Processing*, 47:50–67, 2015.

[20] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.

[21] F. Llorente, L. Martino, D. Delgado, and J. López-Santiago. Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review*, 65(1):3–58, 2023.

[22] L. Martino and V. Elvira. Effective sample size approximations as entropy measures. *Computational Statistics*, 40(9):5433–5464, Dec 2025.

[23] L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[24] L. Martino, D. Luengo, and J. Míguez. Independent random sampling methods. *Springer*, 2018.

[25] J. Mikelson and M. Khammash. Likelihood-free nested sampling for parameter inference of biochemical reaction networks. *PLOS Computational Biology*, 16(10):1–24, 10 2020.

[26] P. Mukherjee, D. Parkinson, and A. R. Liddle. A nested sampling algorithm for cosmological model selection. *The Astrophysical Journal*, 638(2):L51, 2006.

[27] R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–741, 2003.

[28] N. G. Polson and J. G. Scott. Vertical-likelihood Monte Carlo. *arXiv preprint arXiv:1409.3601*, 2014.

[29] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[30] J. Skilling. Nested sampling. *AIP Conference Proceedings*, 735(1):395–405, 11 2004.

[31] J. Skilling. Nested sampling for general Bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.

[32] M. J. Williams, J. Veitch, and C. Messenger. Nested sampling with normalizing flows for gravitational-wave inference. *Physical Review D*, 103(10):103006, 2021.

[33] M. J. Williams, J. Veitch, and C. Messenger. Importance nested sampling with normalising flows. *Mach. Learn. Sci. Tech.*, 4(3):035011, 2023.

[34] Z. Zhao and T. A. Severini. Integrated likelihood computation methods. *Computational Statistics*, 32(1):281–313, 2017.