

Nested Sampling: A Critical and Comprehensive Theoretical Guide

L. Martino^{*†*}, F. Llorente[†],

^{*} Università di Catania, Catania (Italy).

[†] Stony Brook University, New York (USA).

[†] corresponding author: luca.martino@unict.it.

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. The connections between NS, importance sampling (IS), and vertical Monte Carlo are also discussed in detail. In doing so, this work seeks to deepen understanding of the method

^{*}The authors have no conflicts of interest. Any data and code used in this work will be made publicly available upon publication.

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) [45, 46] is a stochastic quadrature method designed to approximate high-dimensional and potentially complex integrals [30, 35]. The main framework of application is Bayesian inference [18, 30, 43]. 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 [5, 11]. Excellent, comprehensive and up-to-date reviews of the methodology and its developments can be found in [1, 5, 26, 28]. 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, 5, 37] and, more recently, also in engineering and statistical signal processing [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 [27, 32, 50]. Indeed, originally NS was introduced precisely for the purpose of computing the marginal likelihood (a.k.a., the Bayesian evidence) [45]. 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 [11]. During this time, numerous efficient implementations, methodological refinements, and diagnostic tools have been proposed to enhance its reliability and

performance [15, 28, 26]. 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 [26].

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 [9, 10, 11, 22]. 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. Another strength of NS is that its theoretical foundations provide clear diagnostic criteria, whose validity, however, relies on the satisfaction of the main assumptions underlying the theory. However, all the NS theory relies on the idea of sampling from *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 [9, 10], 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 [9, 10]. The problem of sampling within domains constrained by likelihood values is, in principle, the same theoretical challenge encountered in *slice sampling* [38]. In both cases, NS

and slice sampling, the development of robust and efficient computational implementations, often incorporating sophisticated internal Monte Carlo (MC) techniques to draw from $g(\boldsymbol{\theta}|\lambda)$, appears to have effectively “solved” the problem, at least from a practical point of view. Another insightful theoretical analysis of NS is presented in [22]. 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 [11, 22, 25].

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. Moreover, we provide a detailed analysis of the connections among nested sampling, importance sampling (IS), and vertical Monte Carlo [41]. We also discuss possible ways of combining these approaches to exploit their complementary strengths. To the best of our knowledge, some aspects of this discussion can constitute novel contributions to the literature.

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, $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_{D_\theta}] \in \Theta \subseteq \mathbb{R}^{D_\theta}$, where $\theta_d \in \mathbb{R}$ for all $d = 1, \dots, D_\theta$, given a set of observed measurements, $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$. We assume that $\Theta \subseteq \mathbb{R}^{D_\theta}$ is a connected parameter space. In a standard Bayesian framework, we assume a known observation 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.,

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

where

$$Z_{\text{ev}} = p(\mathbf{y}) = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\Theta} \pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \quad (2)$$

is the so-called marginal likelihood, a.k.a., Bayesian *evidence*. This quantity is important for model-selection purposes, as we show below. However, usually $Z_{\text{ev}} = p(\mathbf{y})$ is unknown and difficult to approximate. We can evaluate the integrand, i.e., unnormalized target (posterior) function, $\pi(\boldsymbol{\theta})$. Clearly, note that $\bar{\pi}(\boldsymbol{\theta}) = \frac{1}{Z_{\text{ev}}}\pi(\boldsymbol{\theta})$. Several methods have been proposed to compute the marginal likelihood Z_{ev} , and virtually all of them rely on importance sampling (IS) identities [8, 7, 17, 32]. In contrast, the nested sampling (NS) algorithm is primarily based on a different class of representations, often referred to as vertical likelihood representations [41]. 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_{ev} . We begin by deriving two equivalent one-dimensional integral formulations of Z_{ev} , 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 and main assumptions

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

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

$$= \int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} \int_0^{\ell(\mathbf{y}|\boldsymbol{\theta})} d\lambda, \quad (\text{extended space representation}) \tag{4}$$

$$= \int_{\Theta} g(\boldsymbol{\theta})d\boldsymbol{\theta} \int_0^{\infty} \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}d\lambda, \tag{5}$$

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

$$\begin{aligned}
Z_{\text{ev}} &= \int_0^\infty d\lambda \int_{\Theta} g(\boldsymbol{\theta}) \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\} d\boldsymbol{\theta} \\
&= \int_0^\infty d\lambda \int_{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda} g(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int_0^\infty Z(\lambda) d\lambda, \\
&= \int_0^{\sup_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{y}|\boldsymbol{\theta})} Z(\lambda) d\lambda,
\end{aligned} \tag{6}$$

where we have set

$$Z(\lambda) = \int_{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda} g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda), \quad \text{where } \boldsymbol{\theta} \sim g(\boldsymbol{\theta}). \tag{7}$$

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

Assumptions. For simplicity of the exposition, we assume throughout that:

- The likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta})$ is continuous and continuously differentiable with respect to $\boldsymbol{\theta}$ on Θ , where Θ is assumed to be a connected subset of $\mathbb{R}^{D_\theta}$.
- The likelihood function is always strictly positive, $\ell(\mathbf{y}|\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta$.
- The likelihood function $\ell(\mathbf{y}|\boldsymbol{\theta})$ is not constant on Θ . Thus, we avoid “plateaus”, which would complicate the subsequent derivations. For further details, see [16].
- The prior density $g(\boldsymbol{\theta})$ is strictly positive everywhere on Θ , i.e., $g(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta$.
- If $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$, then the induced random variable $L = \ell(\mathbf{y}|\boldsymbol{\theta})$ has a continuous distribution.

- In Eq. (6) and throughout the remainder of the work, we additionally assume that $\ell(\mathbf{y}|\boldsymbol{\theta})$ is bounded on Θ . Hence, the upper limit of integration can be taken to be $\sup_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{y}|\boldsymbol{\theta})$, which is finite.

Some of these conditions can be easily relaxed by formulating them as “almost everywhere” statements, meaning that they could not hold on subsets of Θ of measure zero. 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 $a = Z(\lambda)$ (see Figure 2 for two graphical examples). The generalized inverse function is then:

$$\Lambda(a) = Z^{-1}(a) = \inf\{\lambda \geq 0 : Z(\lambda) \leq a\} \text{ with } a \in [0, 1], \quad (8)$$

and is also non-increasing. Under the above definition and assumptions, note that $Z(\lambda) > a$ if and only if $\lambda < \Lambda(a)$, i.e., $Z(\lambda)$. Moreover, we have the two extreme values:

$$\Lambda(0) = \sup_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{y}|\boldsymbol{\theta}), \quad \Lambda(1) = 0.$$

i.e., $\Lambda(a) : [0, 1] \rightarrow [0, \sup \ell(\mathbf{y}|\boldsymbol{\theta})]$. Moreover, under the above assumptions, $Z(\lambda)$ has neither jumps nor flat regions over the interior of its effective range, $\lambda \in [0, \sup \ell(\mathbf{y}|\boldsymbol{\theta})]$. Then, we can

write

$$\begin{aligned}
Z_{\text{ev}} &= \int_0^\infty Z(\lambda) d\lambda \\
&= \int_0^\infty d\lambda \int_0^1 \mathbb{I}\{a < Z(\lambda)\} da && \text{(again the extended space “trick”)} \\
&= \int_0^1 da \int_0^\infty \mathbb{I}\{a < Z(\lambda)\} d\lambda && \text{(switching the integration order)} \\
&= \int_0^1 da \int_0^\infty \mathbb{I}\{\lambda < \Lambda(a)\} d\lambda && \text{(using } Z(\lambda) > a \iff \lambda < \Lambda(a)) \\
&= \int_0^1 \Lambda(a) da.
\end{aligned} \tag{9}$$

Remark 2. Hence, we have obtained two of the marginal likelihood Z_{ev} in Eqs. (6)-(9), that is generally defined by an highly-multidimensional integral given in Eq. (3).

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 Discussion on possible quadratures

We have thus obtained two alternative one-dimensional integral representations of the Bayesian evidence Z_{ev} , which was originally defined as a multidimensional integral over the parameter space Θ , i.e.,

$$Z_{\text{ev}} = \int_0^{\sup \ell(\mathbf{y}|\boldsymbol{\theta})} Z(\lambda) d\lambda = \int_0^1 \Lambda(a) da. \tag{10}$$

The two integrand functions $a = Z(\lambda)$ and $\lambda = \Lambda(a)$ are graphically represented in Figure 1. Hence, the Bayesian evidence Z_{ev} , originally defined as a multidimensional integral over Θ , admits an equivalent representation as a one-dimensional integral over a bounded domain. This reformulation can substantially simplify its numerical evaluation, and 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:

$$\hat{Z}_{\text{ev}} = \sum_{i=1}^I (\lambda_i - \lambda_{i-1}) Z(\lambda_i), \quad \text{or} \quad (11)$$

$$\hat{Z}_{\text{ev}} = \sum_{i=1}^I (a_i - a_{i-1}) \Lambda(a_i). \quad (12)$$

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 $\pi(\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})$. The behavior of the integrands at both boundaries (but especially near the maximum value of the likelihood function) contributes significantly to the value of the integral. Hence the quadrature approximation can have very bad performance, unless the grid of points is chosen with extreme care. For instance, $Z(\lambda)$ can be so sharply peaked that standard quadrature methods struggle to accurately evaluate the integral, even though it is only one-dimensional. Nested sampling (NS) can be viewed as a stochastic method that “auto-magically” constructs an effective grid for this quadrature. In the remainder of this work, we describe in detail the key concepts, additional variables, procedures, and results required for a proper introduction to the NS method. Table 1 summarizes the main notation of the work.

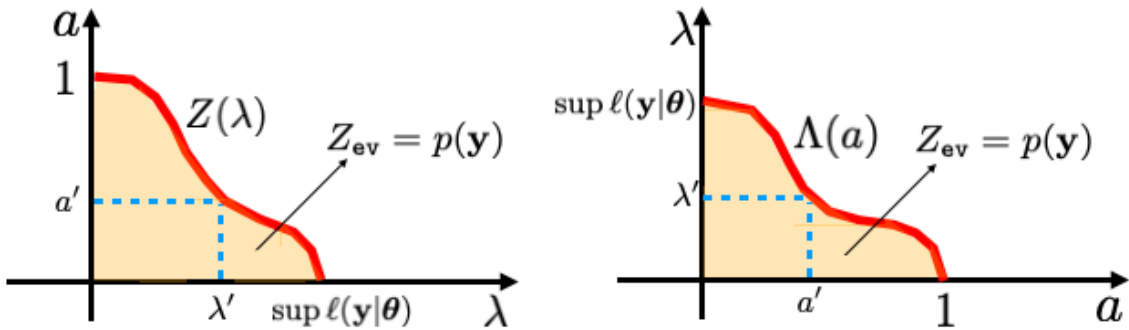


Figure 1: Graphical examples of the two one-dimensional functions $Z(\lambda)$ and $\Lambda(a)$. The area below the curve is in both cases the marginal likelihood Z_{ev} . The marginal likelihood is generally expressed by a multi-dimensional integral, i.e., $Z_{\text{ev}} = \int_{\Theta} \ell(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\the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

Table 1: Main notation of the work.

$\ell(\mathbf{y} \boldsymbol{\theta})$	likelihood function
$g(\boldsymbol{\theta})$	prior density
$\pi(\boldsymbol{\theta}) = \ell(\mathbf{y} \boldsymbol{\theta})g(\boldsymbol{\theta})$	unnormalized posterior function
$\bar{\pi}(\boldsymbol{\theta})$	posterior density, $\bar{\pi}(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta})}{Z_{\text{ev}}}$
Z_{ev}	Bayesian evidence - marginal likelihood
a	Normalized area within $[0, 1]$
λ	likelihood value (ordinate) and threshold value in $g(\boldsymbol{\theta} \lambda)$
$a = Z(\lambda)$	Constrained prior mass, $\int_{\ell(\mathbf{y} \boldsymbol{\theta}) > \lambda} g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{P}(\ell(\mathbf{y} \boldsymbol{\theta}) > \lambda), \boldsymbol{\theta} \sim g(\boldsymbol{\theta})$
$\lambda = \Lambda(a)$	inverse of $Z(\lambda)$, i.e., $Z^{-1}(a)$
L	random variable $L = \ell(\mathbf{y} \boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$
$F(\lambda)$	cumulative function of L , i.e., $F(\lambda) = 1 - Z(\lambda) = \mathbb{P}(\ell(\mathbf{y} \boldsymbol{\theta}) \leq \lambda), \boldsymbol{\theta} \sim g(\boldsymbol{\theta})$
$\tilde{a}$	ratio of normalized areas, e.g., $\tilde{a}_{\text{max}}^{(i)} = \frac{a_{\text{max}}^{(i)}}{a_{\text{max}}^{(i-1)}}$, with i is an iteration index.
$g(\boldsymbol{\theta} \lambda)$	constrained prior pdf, defined as $\frac{1}{Z(\lambda)} \mathbb{I}\{\ell(\mathbf{y} \boldsymbol{\theta}) > \lambda\} g(\boldsymbol{\theta})$
$\mathcal{P}_i = \{\boldsymbol{\theta}_n\}_{n=1}^N$	population of live points, such that $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta} \lambda_{\text{min}}^{(i)})$ with $\boldsymbol{\theta}_n \in \mathcal{P}_i$
$\boldsymbol{\theta}_{\text{rem}}^{(i)}$	removed point, defined as $\arg \min_{\boldsymbol{\theta} \in \mathcal{P}_{i-1}} \ell(\mathbf{y} \boldsymbol{\theta})$, then we set $\mathcal{P}_i = \mathcal{P}_{i-1} \setminus \{\boldsymbol{\theta}_{\text{rem}}^{(i)}\}$.
γ_i	NS quadrature weight $\gamma_i = a_{\text{max}}^{(i-1)} - a_{\text{max}}^{(i)} \in (0, 1)$ and $\gamma_i \approx e^{-\frac{i-1}{N}} - e^{-\frac{i}{N}}$

4 Additional theoretical foundations of NS

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

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

Note also that

$$\boxed{Z(\lambda) = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda), \quad \text{where} \quad \boldsymbol{\theta} \sim g(\boldsymbol{\theta}).} \quad (13)$$

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

$$\boxed{F(\lambda) = 1 - Z(\lambda) = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) \leq \lambda) = \mathbb{P}(L \leq \lambda), \quad \text{with} \quad \boldsymbol{\theta} \sim g(\boldsymbol{\theta}),} \quad (14)$$

is the cumulative function of a random variable $L = \ell(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$ [35, 43] (recall we assume the continuity of the distribution of the random variable L). Note that $F(0) = 0$ and $F(\lambda') = 1$ for all $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$. Hence, $F(\lambda)$ is a cumulative function and $Z(\lambda) = 1 - F(\lambda)$ is a survival function, both representing normalized areas, i.e., probabilities.

4.2 Sampling from the cumulative functions $F(\lambda)$ and $Z(\lambda)$

The following procedure generates samples λ_n from $F(\lambda)$ in Eq. (14):

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

Recalling the assumption of continuity of the distribution of random variable $L = \ell(\mathbf{y}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$, and recalling the inversion method [35, Chapter 2], we can write that the corresponding value $b_n = F(\lambda_n)$ is uniformly distributed in $[0, 1]$ (since F is the cumulative function of the random variable L , and λ_n is a realization of this random variable), i.e.,

$$b_n = F(\lambda_n) \sim \mathcal{U}([0, 1]). \quad (15)$$

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

$$a_n = Z(\lambda_n) = 1 - F(\lambda_n) \sim \mathcal{U}([0, 1]). \quad (16)$$

In summary, finally we have that

if $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta})$, and $\lambda_n = \ell(\mathbf{y} \boldsymbol{\theta}_n) \sim F(\lambda) = 1 - Z(\lambda)$, then $a_n = Z(\lambda_n) \sim \mathcal{U}([0, 1])$.	(17)
--	------

All these expressions rely on the continuity of the distribution of the random variable L . Moreover, note that we can draw samples from $F(\lambda)$, but we cannot evaluate $F(\lambda)$ (or $Z(\lambda)$) because its analytical form is unknown.

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.,

$$g(\boldsymbol{\theta}|\lambda) = \frac{1}{Z(\lambda)} \mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\} g(\boldsymbol{\theta}) = \begin{cases} \frac{g(\boldsymbol{\theta})}{Z(\lambda)}, & \text{if } \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda, \\ 0, & \text{otherwise,} \end{cases} \quad \text{with } 0 < \lambda < \lambda_{\max}, \quad (18)$$

where $\lambda_{\max} = \sup \ell(\mathbf{y}|\boldsymbol{\theta})$. Since we have assumed Θ compact, ℓ continuous, and assuming now that the likelihood admits a unique global maximum, we can define the maximum likelihood (ML) estimator as $\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max \ell(\mathbf{y}|\boldsymbol{\theta})$, the pdf $g(\boldsymbol{\theta}|\lambda)$ becomes increasingly concentrated around $\hat{\boldsymbol{\theta}}_{\text{ML}}$ as λ approaches $\lambda_{\max}$. More precisely, under the above regularity assumptions, the probability measures induced by $g(\boldsymbol{\theta}|\lambda)$ converge weakly to the Dirac probability measure concentrated at the ML estimator, i.e., $g(\boldsymbol{\theta}|\lambda) \xrightarrow{w} \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ML}})$ as λ approaches $\lambda_{\max}$ from below, where $\delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ML}})$ denotes the Dirac probability measure concentrated at the atom $\hat{\boldsymbol{\theta}}_{\text{ML}}$, assigning probability one to $\hat{\boldsymbol{\theta}}_{\text{ML}}$. Hence, with the previous assumptions and recalling that we are assuming $\ell(\mathbf{y}|\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta$ (so that $Z(0) = 1$ and $\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > 0\} = 1$), we have the two extreme cases:

$$g(\boldsymbol{\theta}|0) = g(\boldsymbol{\theta}), \quad (19)$$

$$g(\boldsymbol{\theta}|\lambda) \xrightarrow{w} \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ML}}) \quad \text{with} \quad \lambda \uparrow \lambda_{\max} = \ell(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{ML}}), \quad (20)$$

where $\lambda \uparrow \lambda_{\max}$ means that λ approaches $\lambda_{\max}$ from below. 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

unnormalized truncated prior, i.e., $\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}g(\boldsymbol{\theta})$. To normalize it, we have to divide by the value $Z(\lambda)$ (that is represented by the green area), i.e., obtaining $g(\boldsymbol{\theta}|\lambda) = \frac{1}{Z(\lambda)}\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}g(\boldsymbol{\theta})$.

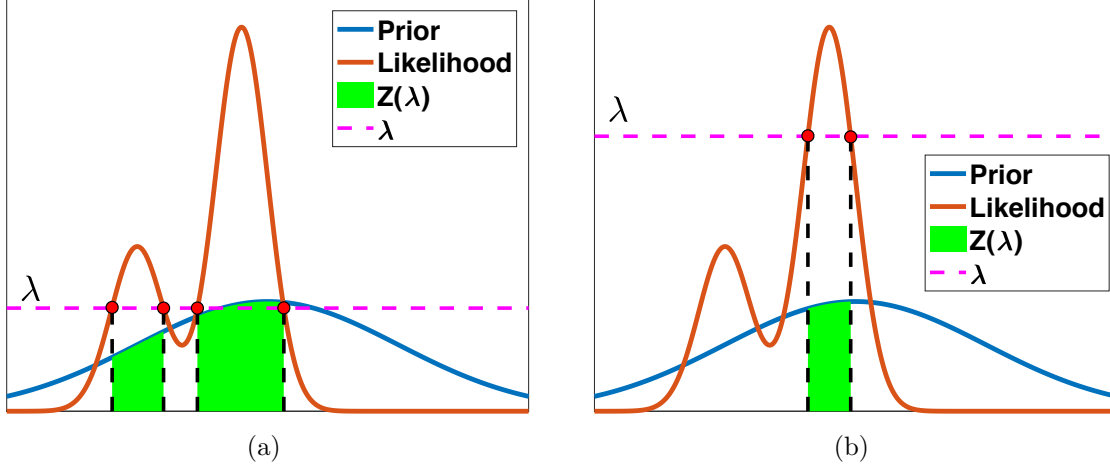


Figure 2: Two examples of the area below the *non-normalized* truncated prior $\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}g(\boldsymbol{\theta})$. The green area represents the function $Z(\lambda)$. The normalized truncated prior $g(\boldsymbol{\theta}|\lambda) = \frac{1}{Z(\lambda)}\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}g(\boldsymbol{\theta})$ is given in Eq. (18). 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 the non-normalized truncated prior $\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}g(\boldsymbol{\theta})$; for obtaining the normalized version $g(\boldsymbol{\theta}|\lambda)$ we have to multiply by $1/Z(\lambda)$. 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 computationally complex).*

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

Given a fixed value $0 < \lambda_0 < \lambda_{\max}$, in order to generate samples from $g(\boldsymbol{\theta}|\lambda_0)$ one possibility is to use an MCMC procedure. However, as an example, the following acceptance-rejection procedure can be also employed [35]:

1. For $n = 1, \dots, N$:
 - (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 $\boldsymbol{\theta}_{n|0} \sim g(\boldsymbol{\theta}|\lambda_0)$ 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 $n = 1, \dots, N$, and the probability of accepting a generated sample $\boldsymbol{\theta}'$ is exactly $Z(\lambda_0) = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda_0)$ with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$, by definition. The values $\lambda_{n|0} = \ell(\mathbf{y}|\boldsymbol{\theta}_{n|0})$ where $\boldsymbol{\theta}_{n|0} \sim g(\boldsymbol{\theta}|\lambda_0)$, have the following *truncated* cumulative function:

$$F(\lambda|\lambda_0) = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) \leq \lambda | \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda_0) = \frac{F(\lambda) - F(\lambda_0)}{1 - F(\lambda_0)}, \quad \text{with } \lambda \geq \lambda_0, \text{ and } \boldsymbol{\theta} \sim g(\boldsymbol{\theta}|\lambda_0), \quad (21)$$

i.e., we can write $\lambda_n \sim F(\lambda|\lambda_0)$. Recall that $F(\lambda) = \mathbb{P}(\ell(\mathbf{y}|\boldsymbol{\theta}) \leq \lambda) = \mathbb{P}(L \leq \lambda)$, with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$ and we have assumed the continuity of distribution of L . Hence, considering the steps above, (a) draw $\boldsymbol{\theta}' \sim g(\boldsymbol{\theta})$ and (b) if $\ell(\mathbf{y}|\boldsymbol{\theta}') > \lambda_0$ then set $\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}')$ (otherwise reject $\boldsymbol{\theta}'$ and repeat), 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. (14). 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 $\lambda' \geq \sup \ell(\mathbf{y}|\boldsymbol{\theta})$.

Remark 4. Consider the accept-reject procedure described above and let $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ be the N accepted samples, drawn independently from $g(\boldsymbol{\theta})$ and satisfying $\ell(\mathbf{y}|\boldsymbol{\theta}_n) > \lambda_0$ for $n = 1, \dots, N$. Consequently, the population $\mathcal{P} = \{\boldsymbol{\theta}_n\}_{n=1}^N$ consists of independent samples from $g(\boldsymbol{\theta}|\lambda_0)$. Now, let $\lambda' > \lambda_0$ be a fixed, nonrandom threshold and consider the event that all N samples satisfy the stricter constraint $\ell(\mathbf{y}|\boldsymbol{\theta}_n) > \lambda'$. Conditional on this event, the samples remain independent and are distributed according to $g(\boldsymbol{\theta}|\lambda')$.

Remark 5. A key observation is the following. Consider again the N accepted samples $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$, drawn independently from $g(\boldsymbol{\theta})$ and satisfying $\ell(\mathbf{y}|\boldsymbol{\theta}_n) > \lambda_0$ for all $n = 1, \dots, N$. Define $\lambda_{\min} = \min_{n=1, \dots, N} \ell(\mathbf{y}|\boldsymbol{\theta}_n)$ and let $\boldsymbol{\theta}_{n'}$ denote the sample attaining this minimum, so that $\lambda_{\min} = \ell(\mathbf{y}|\boldsymbol{\theta}_{n'}) > \lambda_0$. After removing $\boldsymbol{\theta}_{n'}$, the remaining $N - 1$ samples can be reused as independent draws from $g(\boldsymbol{\theta}|\lambda_{\min})$, conditionally on $\lambda_{\min}$ (see also Figure 5), provided that the likelihood values have a continuous distribution, so that ties occur with probability zero.

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. (17), we can write:

if $\boldsymbol{\theta}_{n 0} \sim g(\boldsymbol{\theta} \lambda_0)$, and $\lambda_{n 0} = \ell(\mathbf{y} \boldsymbol{\theta}_{n 0}) \sim F(\lambda \lambda_0)$, then $a_{n 0} = Z(\lambda_{n 0}) \sim \mathcal{U}([0, a_0])$, with $a_0 = Z(\lambda_0) \leq 1$.	(22)
---	------

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:

$\tilde{a}_n = \frac{a_{n 0}}{a_0} = \frac{Z(\lambda_{n 0})}{Z(\lambda_0)} \sim \mathcal{U}([0, 1]), \quad \forall n = 1, \dots, N.$	(23)
--	------

Remark 6. If we can draw samples $\boldsymbol{\theta}_{n|0}$ exactly from $g(\boldsymbol{\theta}|\lambda_0)$ and evaluating $\lambda_{n|0} = \ell(\mathbf{y}|\boldsymbol{\theta}_{n|0})$, the ratio of areas $\frac{a_{n|0}}{a_0} = \frac{Z(\lambda_{n|0})}{Z(\lambda_0)}$ is uniformly distributed on $[0, 1]$ for every $\lambda_0 \in (0, \lambda_{\max})$. This is a powerful result, as it holds regardless of the specific forms of the likelihood and prior functions and for any (admissible) values of λ_0 and $\lambda_{n|0}$. This theoretical result, which relies on the ability to sample exactly from $g(\boldsymbol{\theta}|\lambda_0)$, is a key property of the NS procedure and provides the foundation

for the subsequent theoretical developments.

4.6 Distribution of $\tilde{a}_{\max}$

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

$$\lambda_{\min|0} = \min_n \lambda_{n|0}, \quad a_{\max|0} = Z(\lambda_{\min|0}), \quad \text{and} \quad \tilde{a}_{\max} = \frac{a_{\max|0}}{a_0} = \frac{Z(\lambda_{\min|0})}{Z(\lambda_0)}. \quad (24)$$

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

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

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

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

is distributed according to a Beta distribution $\mathcal{B}(N, 1)$, i.e., with cumulative function $F_{\max}(\tilde{a}) = \tilde{a}^N$ and density $f_{\max}(\tilde{a}) = \frac{dF_{\max}(\tilde{a})}{d\tilde{a}} = N\tilde{a}^{N-1}$ [35, Section 2.3.6]. In summary, we have

$$\tilde{a}_{\max} = \frac{Z(\lambda_{\min|0})}{Z(\lambda_0)} \sim \mathcal{B}(N, 1),$$

where $\lambda_{\min|0} = \min_n \lambda_{n|0},$

and $\lambda_{1|0}, \dots, \lambda_{N|0} \sim F(\lambda|\lambda_0).$

(25)

It is important to remark that all these expressions from (22) to (25) rely on the continuity of the distribution of the random variable $L = \ell(\mathbf{y}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$. 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 2.

Table 2: 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 $\lambda_{n 0} = \ell(\mathbf{y} \boldsymbol{\theta}_{n 0}) \sim F(\lambda \lambda_0)$, $\tilde{a}_n = \frac{a_{n 0}}{a_0} = \frac{Z(\lambda_{n 0})}{Z(\lambda_0)} \sim \mathcal{U}([0, 1]).$
4.6	If $\boldsymbol{\theta}_{n 0} \sim g(\boldsymbol{\theta} \lambda_0)$, and $\lambda_{n 0} = \ell(\mathbf{y} \boldsymbol{\theta}_{n 0}) \sim F(\lambda \lambda_0)$, $n = 1, \dots, N$, $\lambda_{\min 0} = \min_n \lambda_{n 0},$ $\tilde{a}_{\max} = \max \tilde{a}_n = \frac{Z(\lambda_{\min 0})}{Z(\lambda_0)} \sim \mathcal{B}(N, 1).$

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 (10) [46, 11, 41]. NS estimates Z_{ev} by a quadrature using nodes (in *decreasing* order),

$$0 < a_{\max}^{(I)} < \dots < a_{\max}^{(1)} < a_{\max}^{(0)} = 1,$$

and the rectangular quadrature formula:

$$\hat{Z}_{\text{ev}} = \sum_{i=1}^I (a_{\max}^{(i-1)} - a_{\max}^{(i)}) \Lambda(a_{\max}^{(i)}) = \sum_{i=1}^I (a_{\max}^{(i-1)} - a_{\max}^{(i)}) \lambda_{\min}^{(i)}, \quad (26)$$

with $a_{\max}^{(0)} = 1$ and recall that we always have $a_{\max}^{(i-1)} - a_{\max}^{(i)} > 0$. Furthermore, we will see that the NS construction yields that

$$\boxed{\lim_{I \rightarrow \infty} a_{\max}^{(I)} = a_{\max}^{(\infty)} = 0.} \quad (27)$$

See Figure 3 for a graphical representation.

Remark 7. Recall that $\lambda = \ell(\mathbf{y}|\boldsymbol{\theta})$'s represent likelihood values, and a 's represent normalized area values, hence contained in $[0, 1]$.

In a standard quadrature rule, we should to specify the grid points $a_{\max}^{(i)}$'s (possibly well-located, with a suitable strategy) and the corresponding values $\lambda_{\min}^{(i)} = \Lambda(a_{\max}^{(i)})$. However, the NS algorithm automatically selects the values $\lambda_{\min}^{(i)}$ and approximates the corresponding values $a_{\max}^{(i)}$. Recall that the function $\Lambda(a)$, and its inverse $a = \Lambda^{-1}(\lambda) = Z(\lambda)$, are generally intractable and unknown, so that it is not even possible to evaluate $\Lambda(a)$ at a grid of chosen $a_{\max}^{(i)}$'s.

Remark 8. The quadrature values $a_{\max}^{(i)}$, $\lambda_{\min}^{(i)}$ in Eq. (26), are automatically chosen by the NS procedure (see Figure 3). The NS algorithm first suitably selects the ordinates $\lambda_{\min}^{(i)}$ (likelihood values) with a stochastic process that accumulates more values close to $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$, and then uses approximations $\hat{a}_i$ of the corresponding (unknown) values $a_{\max}^{(i)} = Z(\lambda_{\min}^{(i)})$. This last step is possible since the distribution of $a_{\max}^{(i)}$ is known under the given assumptions (see Section 4.6).

Clearly, alternative quadrature rules could be also employed. For instance, the trapezoidal quadrature formula for NS is:

$$\hat{Z}_{\text{ev}} = \sum_{i=1}^I \frac{\Lambda(a_{\max}^{(i-1)}) + \Lambda(a_{\max}^{(i)})}{2} (a_{\max}^{(i-1)} - a_{\max}^{(i)}).$$

Defining $\gamma_i = a_{\max}^{(i-1)} - a_{\max}^{(i)}$, in each interval $[a_{\max}^{(i-1)}, a_{\max}^{(i)}]$, bounds for the quadrature error of the

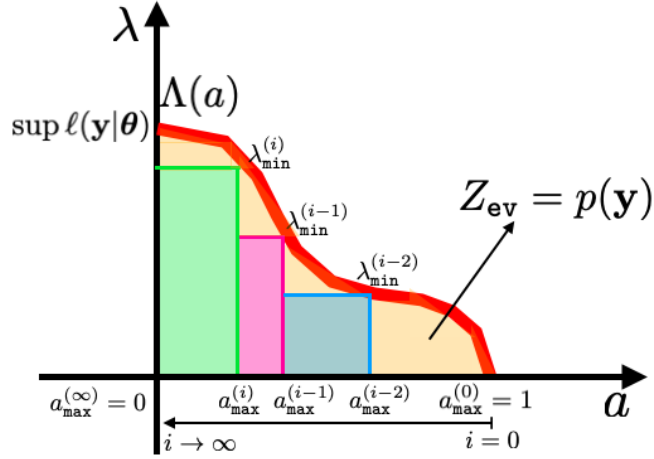


Figure 3: Graphical representation of the adaptive quadrature scheme provided by the NS procedure, where the nodes $(a_{\max}^{(i)}, \lambda_{\min}^{(i)})$, are automatically chosen by the algorithm. More precisely, in practice, the algorithm first chooses the value $\lambda_{\min}^{(i)} = \Lambda(a_{\max}^{(i)})$ and then estimate/approximate the corresponding value $a_{\max}^{(i)}$, with $\hat{a}_i \approx a_{\max}^{(i)}$ as shown in Eq. (35).

rectangular and trapezoidal formulas can be found:

$$\left| E_{\text{rect}}^{(i)} \right| \leq \frac{1}{2} \sup |\Lambda'(\xi)| \gamma_i^2, \quad \left| E_{\text{trap}}^{(i)} \right| \leq \frac{1}{12} \sup |\Lambda''(\xi)| \gamma_i^3, \quad (28)$$

where $\xi \in [a_{\max}^{(i)}, a_{\max}^{(i-1)}]$ and $\Lambda'(\xi)$, $\Lambda''(\xi)$ denote the first and second derivatives of $\lambda(a)$, respectively, evaluated at ξ . However, in the NS procedure, the prior areas $a_{\max}^{(i)}$ are estimated, and the likelihood levels $\lambda_{\min}^{(i)} = \Lambda(a_{\max}^{(i)})$ are random order statistics. Consequently, the error formulas above describe only a partial error of the NS algorithm (i.e., only the quadrature component). The complete NS error of the evidence estimator must include the statistical error components of the overall procedure. See Section 5.5, for some other details.

5.2 Choice of $\lambda_{\min}^{(i)}$ and $a_{\max}^{(i)}$ 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 $\lambda_{\min}^{(i)}$, $i = 1, \dots, I$, such that

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

Table 3 provides a compact and complete description of the standard NS algorithm that is based on the sampling of the truncated prior pdf $g(\boldsymbol{\theta}|\lambda_{\min}^{(i-1)})$ (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 $\lambda_{\min}^{(0)} = 0$ and $a_{\max}^{(0)} = Z(\lambda_{\min}^{(0)}) = 1$. Then, N samples are drawn from the prior $\boldsymbol{\theta}_n \sim g(\boldsymbol{\theta}|\lambda_{\min}^{(0)}) = g(\boldsymbol{\theta})$ obtaining an (initial) cloud

$$\mathcal{P}_0 = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N\} = \{\boldsymbol{\theta}_n\}_{n=1}^N, \quad (30)$$

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

$$\lambda_{\min}^{(1)} = \min_n \lambda_n = \min_{\boldsymbol{\theta} \in \mathcal{P}_0} \ell(\mathbf{y}|\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta}_{\text{rem}}^{(1)}),$$

where $\boldsymbol{\theta}_{\text{rem}}^{(1)} = \arg \min_{\boldsymbol{\theta} \in \mathcal{P}_0} \ell(\mathbf{y}|\boldsymbol{\theta})$ is also removed from $\mathcal{P}_0$, i.e., $\mathcal{P}_1 = \mathcal{P}_0 \setminus \{\boldsymbol{\theta}_{\text{rem}}^{(1)}\}$ (at this stage, now we have $|\mathcal{P}_1| = N - 1$). Moreover, since $\{\lambda_n\}_{n=1}^N \sim F(\lambda)$, using the result in Eq. (25), we have that

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

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

- At a generic i -th iteration, with $i \geq 2$, a unique additional sample $\boldsymbol{\theta}'$ is drawn from the truncated prior $g(\boldsymbol{\theta}|\lambda_{\min}^{(i-1)})$ and added to the current cloud of samples, i.e., $\mathcal{P}_{i-1} = \mathcal{P}_{i-1} \cup \boldsymbol{\theta}'$ (now we recover $|\mathcal{P}_{i-1}| = N$). First of all, note that the new value $\lambda' = \ell(\mathbf{y}|\boldsymbol{\theta}')$ is

distributed as $F(\lambda|\lambda_{\min}^{(i-1)})$ by construction (see Section 4.3). More precisely, all the N ordinate (likelihood) values:

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

are distributed as $F(\lambda|\lambda_{\min}^{(i-1)})$, i.e., $\{\lambda_n\}_{n=1}^N \sim F(\lambda|\lambda_{\min}^{(i-1)})$. This is due to how the population $\mathcal{P}_{i-1}$ has been built in the previous iterations. For the same reason, at this stage, all the N samples $\boldsymbol{\theta}_n \in \mathcal{P}_{i-1}$ are distributed as $g(\boldsymbol{\theta}|\lambda_{\min}^{(i-1)})$, as stated in Remarks 4 and 5. Then, we compute again the new minimum value as

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

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

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

where we have used Eq. (25). Note that we have also found the recursion among the following random variables,

$$a_{\max}^{(i)} = \tilde{a}_{\max}^{(i)} a_{\max}^{(i-1)}, \quad (32)$$

for $i = 1, \dots, I$, recalling that $a_{\max}^{(0)} = 1$.

- We do not know the actual values of all the areas $a_{\max}^{(i)}$. The idea is to estimate the random values $a_{\max}^{(i)}$ (and $\tilde{a}_{\max}^{(i)}$) for all i , by their expected values. With this purpose, we have to observe that:

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

Beta distribution $\mathcal{B}(N, 1)$, i.e., $\tilde{a}_{\max}^{(i)} \approx \mathbb{E}[\tilde{a}_{\max}^{(i)}]$, where

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

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 $\frac{N}{N+1}$ as N grows, as depicted in Figure 4 (this is already a very good approximation for $N \geq 7$). Another explanation is often given: since $\mathbb{E}[\log \tilde{a}_{\max}^{(i)}] = -\frac{1}{N}$, if we take then the exponential $\exp\left(\mathbb{E}[\log \tilde{a}_{\max}^{(i)}]\right) = \exp\left(-\frac{1}{N}\right)$.¹ In any case, the need of this approximation is obscure since $\mathbb{E}[\tilde{a}_{\max}^{(i)}] = \frac{N}{N+1}$ is analytically available, i.e., the use of $\exp(-1/N)$ is not strictly necessary.

- Additionally we need to assume *the independence of the live points in the population at i -th iteration*, conditional on $\lambda_{\min}^{(i-1)}$. Under this assumption, in Eq. (32), we can use the property of the expectation valid for independent variables²,

$$\mathbb{E}[a_{\max}^{(i)}] = \mathbb{E}[\tilde{a}_{\max}^{(i)}] \mathbb{E}[a_{\max}^{(i-1)}], \quad (34)$$

and also using the approximation in (33), the recursion in Eq. (32) becomes:

$$a_{\max}^{(i)} \approx \exp\left(-\frac{1}{N}\right) a_{\max}^{(i-1)} = a_{\max}^{(0)} \prod_{k=1}^i \exp\left(-\frac{1}{N}\right) = \exp\left(-\frac{i}{N}\right), \quad \forall i, \quad (35)$$

where we have also considered that $a_{\max}^{(0)} = 1$. Then, denoting $\hat{a}_i = \exp\left(-\frac{i}{N}\right)$, we can use $\hat{a}_i$ as an approximation of $a_{\max}^{(i)}$, that are the coefficients employed in Eq. (26) and, after the corresponding substitution with $\hat{a}_i$, in Eq. (38). Clearly, note that $\lim_{i \rightarrow \infty} \hat{a}_i = 0$.

¹Clearly, $\exp\left(\mathbb{E}[\log \tilde{a}_{\max}^{(i)}]\right) \leq \mathbb{E}[\tilde{a}_{\max}^{(i)}]$ by Jensen inequality, i.e., $f(\mathbb{E}[A]) \leq \mathbb{E}[f(A)]$ where $f(A) = \exp(A)$ is a convex function, and $A = \log \tilde{a}_{\max}^{(i)}$ (see also Figure 33).

² Assuming the independence of the live points in the population at i -th iteration, the variables $\tilde{a}_{\max}^{(i)}$ and $a_{\max}^{(i-1)}$ are independent, due to Eq. (31), i.e., that ratio has always the same beta distribution, $\mathcal{B}(N, 1)$, for all i .

Remark 9. *It is worth noting that, when rejection sampling is used to sample from $g(\boldsymbol{\theta}|\lambda_{\min}^{(i)})$, the quantity $a_{\max}^{(i)} = Z(\lambda_{\min}^{(i)})$ can be also estimated using the observed acceptance rate, since the theoretical acceptance probability is exactly $Z(\lambda_{\min}^{(i)})$, assuming that the proposal distribution is the prior $g(\boldsymbol{\theta})$.*

The resulting algorithm is given in Table 3. Figure 5 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.

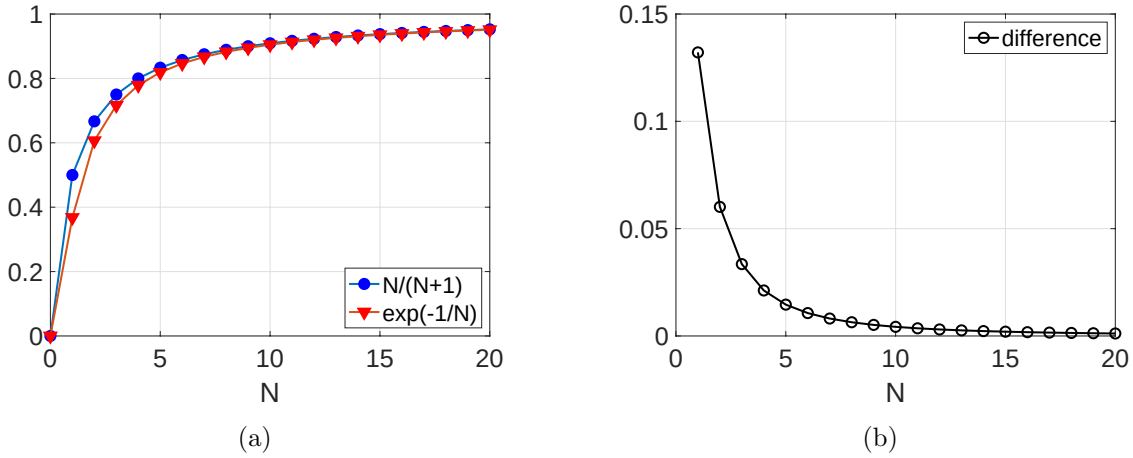


Figure 4: **(a)** The two functions $\frac{N}{N+1}$ and $\exp(-\frac{1}{N})$ as function of the integer N . **(b)** The difference between the two functions (versus N) that converges to zero, thereby justifying Eq. (33).

Remark 10. *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 $\hat{a}_i = \exp(-\frac{i}{N})$ as an approximation of $a_{\max}^{(i)}$.*

Remark 11. *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 $\boldsymbol{\theta}_n \in \mathcal{P}_i$ approaches the value $\sup \ell(\mathbf{y}|\boldsymbol{\theta})$.*

Remark 12. *Note that with respect to the θ -space, the NS method is a population sampler where*

a cloud of points $\mathcal{P}_i = \{\boldsymbol{\theta}_n\}_{n=1}^N$ that changes with the iterations, adding and removing one point per iteration.

Table 3: The standard nested sampling procedure.

1. Choose the number of points per iterations N , the total number of iterations I (big enough, or a stopping rule; see Section 5.3), and set $\hat{a}_0 = 1$.

2. Draw $\{\boldsymbol{\theta}_n\}_{n=1}^N \sim g(\boldsymbol{\theta})$ and define the set $\mathcal{P}_0 = \{\boldsymbol{\theta}_n\}_{n=1}^N$. Let us also define the notation

$$\ell(\mathbf{y}|\boldsymbol{\theta} \in \mathcal{P}_0) = \{\lambda_n = \ell(\mathbf{y}|\boldsymbol{\theta}_n) \text{ for all } \boldsymbol{\theta}_n \in \mathcal{P}_0\}, \quad (36)$$

3. Set $\lambda_{\min}^{(1)} = \min_{\boldsymbol{\theta} \in \mathcal{P}_0} \ell(\mathbf{y}|\boldsymbol{\theta}) = \ell(\mathbf{y}|\boldsymbol{\theta}_{\text{rem}}^{(1)})$ where $\boldsymbol{\theta}_{\text{rem}}^{(1)} = \arg \min_{\boldsymbol{\theta} \in \mathcal{P}_0} \ell(\mathbf{y}|\boldsymbol{\theta})$.

4. Set $\mathcal{P}_1 = \mathcal{P}_0 \setminus \{\boldsymbol{\theta}_{\text{rem}}^{(1)}\}$, i.e., remove $\boldsymbol{\theta}_{\text{rem}}^{(1)}$ from $\mathcal{P}_0$.

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

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

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

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

(c) Set $\mathcal{P}_i = \mathcal{P}_{i-1} \setminus \{\boldsymbol{\theta}_{\text{rem}}^{(i)}\}$, i.e., remove $\boldsymbol{\theta}_{\text{rem}}^{(i)}$ from the set of *live points* $\mathcal{P}_{i-1}$.

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

$$\hat{a}_i = \exp\left(-\frac{i}{N}\right) \approx a_{\max}^{(i)}, \quad (37)$$

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

7. Return

$$\hat{Z}_{\text{ev}} = \sum_{i=1}^I (\hat{a}_{i-1} - \hat{a}_i) \lambda_{\min}^{(i)} = \sum_{i=1}^I (e^{-\frac{i-1}{N}} - e^{-\frac{i}{N}}) \lambda_{\min}^{(i)}. \quad (38)$$

Remark 13. Denoting $\Theta_i = \{\boldsymbol{\theta} \in \Theta : \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda_{\min}^{(i)}\}$ with $\lambda_{\min}^{(0)} = 0 < \lambda_{\min}^{(1)} < \dots < \lambda_{\min}^{(I)}$, then the regions Θ_i are nested, i.e., $\Theta_I \subseteq \Theta_{I-1} \subseteq \dots \subseteq \Theta_0 = \Theta$. Thus, the constrained regions overlap, because each new region is contained in the preceding one. Hence, there is not a partition of the

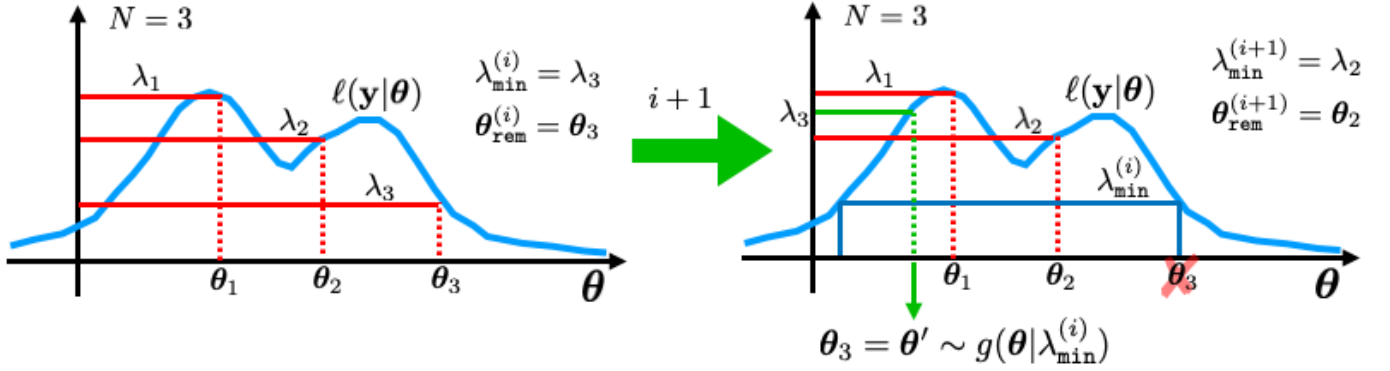


Figure 5: Graphical representation of the NS procedure from the i -th iteration to the $(i+1)$ -th iteration, with $N = 3$ live points within the population $\mathcal{P}_i = \{\theta_1, \theta_2, \theta_3\}$. The blue line represents the likelihood function $\ell(\mathbf{y}|\theta)$. At i -th iteration, the point θ_3 has the lowest likelihood value, i.e., $\lambda_{\min}^{(i)} = \lambda_3 = \ell(\mathbf{y}|\theta_3)$. Hence, θ_3 is removed from the population $\mathcal{P}_i$ and, at the next $(i+1)$ -th iteration, a new sample is generated $\theta_3 = \theta' \sim g(\theta|\lambda_{\min}^{(i)})$ and add to the set $\mathcal{P}_i$ in order to keep fixed the number of live points $N = |\mathcal{P}_i| = 3$. The new sample generation, at the $(i+1)$ -th iteration, is done according to the truncated prior $g(\theta|\lambda_{\min}^{(i)})$ (see Figure 2). New live points tend to be added in regions of high probability mass (and, after several iterations, near to the global mode).

parameter space Θ . However, although the constrained regions in Θ -space overlap, NS constructs disjoint shells by taking subsets defined as the difference $\mathcal{S}_i = \Theta_{i-1} \setminus \Theta_i$. Ignoring boundary aspects, these subsets do not overlap, $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$, $i \neq j$. This is done taking into account the rest $\hat{a}_{i-1} - \hat{a}_i$ (see also Figure 3). Each subset corresponds to a range of likelihood values, and the evidence can be decomposed into contributions from these subsets, plus the final remaining region.

5.3 Stopping rules suggested in literature

The nested sampling in Table 3 is an idealized/asymptotic version of the algorithm where the number of iterations I should be enough, i.e., $I \rightarrow \infty$. A number of iterations I and/or maximum number of likelihood evaluation can be always considered following a computational budget, but they are not a convergence criterion [24, 26]. However, given I finite iteration, the *residual evidence*, i.e., the amount of evidence that has not been included in final estimator of the NS

procedure in Table 3, can be estimated:

$$\boxed{Z_{\text{res}}^{(I)} \approx \int_0^{\hat{a}_I} \Lambda(a) da = \int_0^{\exp(-\frac{I}{N})} \Lambda(a) da.} \quad (39)$$

Moreover, as shown in Figure 1, we have

$$\Lambda(a) \leq \sup_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{y}|\boldsymbol{\theta}) \approx L_{\max}^{(I)} = \max_{\boldsymbol{\theta} \in \mathcal{P}_I} \ell(\mathbf{y}|\boldsymbol{\theta}_n), \quad \boldsymbol{\theta}_n \in \mathcal{P}_I,$$

as a consequence, we have:

$$\boxed{Z_{\text{res}}^{(I)} \approx \hat{Z}_{\text{res}}^{(I)} = \exp\left(-\frac{I}{N}\right) L_{\max}^{(I)},} \quad (40)$$

that is a practical approximation of the residual evidence that can be used for designing a stopping rule [14, 25, 26, 47]. The run is stopped when this becomes sufficiently small relative to the accumulated evidence, for example

$$\frac{\hat{Z}_{\text{res}}^{(I)}}{\hat{Z}_{\text{ev}}^{(I)}} < \epsilon, \quad \text{with, e.g., } \epsilon = 0.01,$$

or the same criterion but in logarithmic domain, this is essentially the criterion used by [14, 47]. An alternative idea is to estimate the variance $\hat{\sigma}_Z$ of $\hat{Z}_{\text{ev}}^{(I)}$, for instance, using parallel runs of the NS procedures, and considering the stopping rule $\frac{\hat{\sigma}_Z}{\hat{Z}_{\text{ev}}^{(I)}} < \epsilon$. A simpler stopping criterion can also be employed, based on the stabilization of the evidence estimate:

$$\left| \log \hat{Z}_{\text{ev}}^{(I)} - \log \hat{Z}_{\text{ev}}^{(I-k)} \right| < \epsilon.$$

for some number k of consecutive iterations. This has an intuitive appeal, but there is an important problem: small changes in the current estimate do not guarantee that the unexplored region contains negligible evidence. Indeed, narrow posterior peaks can be missed or encountered late,

so a criterion based only on the stability of successive estimates can be misleading.

5.4 Refined finite-run NS estimator

The NS estimator in Eq. (38) is based on the contributions of the discarded points and is typically considered in the limit $I \rightarrow \infty$. For a finite number of iterations, the estimator can be refined by also accounting for the contribution of the remaining live points, yielding a finite-run approximation [24, 26].

5.4.1 Contribution of the remaining live points

In the same fashion of the residual evidence estimation, we can refine the estimator $\widehat{Z}_{\text{ev}}$ in Eq. (38) by incorporating the contribution of the *final live points* remaining at the last iteration I . Thus, the NS algorithm in Table 3 stops after I iterations and retains N live points in the population $\mathcal{P}_I$, that we denote here as $\boldsymbol{\theta}_1^{(I)}, \dots, \boldsymbol{\theta}_N^{(I)}$. The residual area estimated by these live points can be expressed as [23, 25, 26, 47]:

$$\widehat{Z}_{\text{live}}^{(I)} = \widehat{a}_I \frac{1}{N} \sum_{n=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)}). \quad (41)$$

5.4.2 IS interpretation of Eq. (41)

The estimator above can be seen as an importance sampling (IS) estimator. Indeed, recall that $a_{\text{max}}^{(I)} = Z(\lambda_{\text{min}}^{(I)})$ is the prior mass below the unnormalized constrained prior $\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda_{\text{min}}^{(I)}\}g(\boldsymbol{\theta})$. If we draw from the following proposal density (recall Remarks 4 and 5):

$$\boldsymbol{\theta}_n^{(I)} \sim g(\boldsymbol{\theta}|\lambda_{\text{min}}^{(I-1)}) = \frac{1}{a_{\text{max}}^{(I-1)}} \mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda_{\text{min}}^{(I)}\}g(\boldsymbol{\theta}), \quad n = 1, \dots, N,$$

the IS estimator of posterior area restricted to the region $\{\boldsymbol{\theta} \in \Theta : \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda_{\min}^{(I)}\}$ is given by:

$$\begin{aligned}
\widehat{Z}_{\text{IS}}^{(I)} &= \frac{1}{N} \sum_{n=1}^N \frac{\ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)})g(\boldsymbol{\theta}_n^{(I)})}{\frac{1}{a_{\max}^{(I)}} \mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)}) > \lambda_{\min}^{(I)}\}g(\boldsymbol{\theta}_n^{(I)})}, \\
&= \frac{1}{N} \sum_{n=1}^N \frac{\ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)})}{\frac{1}{a_{\max}^{(I)}}} = \frac{1}{N} \sum_{n=1}^N w_n^{(I)}, \\
&= a_{\max}^{(I)} \frac{1}{N} \sum_{n=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)}) \approx \widehat{a}_I \frac{1}{N} \sum_{n=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)}),
\end{aligned} \tag{42}$$

that is exactly $\widehat{Z}_{\text{live}}^{(I)}$ above.

Remark 14. Note that we have denoted as $w_n^{(I)} = a_{\max}^{(I)} \ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)})$ the posterior IS weights of the remaining live points. Indeed, in Eq. 42, the arithmetic mean of $w_n^{(I)}$ provides the area estimation given by the remaining live points at the last iteration I .

5.4.3 Refined NS estimator

We can include the additional contribution of the remaining live point to the final NS estimator:

$$\begin{aligned}
\widehat{Z}_{\text{ev}} &= \sum_{i=1}^I (\widehat{a}_{i-1} - \widehat{a}_i) \lambda_{\min}^{(i)} + \widehat{Z}_{\text{live}}^{(I)}, \\
\boxed{\widehat{Z}_{\text{ev}} &= \sum_{i=1}^I (\widehat{a}_{i-1} - \widehat{a}_i) \lambda_{\min}^{(i)} + \widehat{a}_I \frac{1}{N} \sum_{n=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)})}.}
\end{aligned} \tag{43}$$

5.5 Main sources of uncertainty in NS

The quadrature rule considered in Section 5.1 introduces a first error component in the final NS estimator. Here, we describe other sources of uncertainty in the NS procedure [11, 22, 25, 46]. A complete list of the main error components in the NS algorithm is given below:

- **Numerical quadrature:** due to the analytic form of the NS estimator, that is discussed in Section 5.1;

- **Constrained-prior mass estimations:** due to the estimation of constrained prior areas $a_{\max}^{(i)}$, which are random realizations. This is discussed in Section 5.5.1 below.
- **Ordinary Monte Carlo variability:** this component is generated to the random variation in likelihood evaluations and live-point configurations; see Section 5.5.2 below.
- **Imperfect sampling constrained-prior:** this can yield potential systematic error due to incorrect or insufficient exploration of constrained regions. It is also described in Section 5.5.2 below.

5.5.1 Uncertainty in the constrained prior area estimations

The constrained prior masses $a_{\max}^{(i)}$ are random variables and we estimate them using their expected value and, with other approximations, we arrive to $\hat{a}_i = \exp\left(-\frac{i}{N}\right) \approx a_{\max}^{(i)}$. Since in the standard NS algorithm, the approximation $\exp\left(-\frac{1}{N}\right)$ of $\frac{N}{N+1}$ is used, and since $\exp\left(\mathbb{E}[\log \tilde{a}_{\max}^{(i)}]\right) = \exp\left(-\frac{1}{N}\right)$, we work here in the logarithmic domain, where we have:

$$\mathbb{E}[\log \tilde{a}_{\max}^{(i)}] = -\frac{1}{N}, \quad \text{Var} [\log \tilde{a}_{\max}^{(i)}] = \frac{1}{N^2}.$$

Consequently, following the recursion in Eq. (32), we also have:

$$\mathbb{E}[\log a_{\max}^{(i)}] = -\frac{i}{N}, \quad \text{Var} [\log a_{\max}^{(i)}] = \frac{i}{N^2},$$

starting with $a_{\max}^{(0)} = 1$. Thus, the uncertainty in the logarithm of the $a_{\max}^{(i)}$ is often defined by the formula $\text{std} [\log a_{\max}^{(i)}] = \frac{\sqrt{i}}{N}$ [25, 46].

5.5.2 Stochastic uncertainty under exact sampling and possible bias

Even if every replacement point is drawn exactly from the suitable constrained prior, the evidence estimate remains random for two reasons:

- **Prior-mass uncertainty:** the ratios $\tilde{a}_{\max}^{(i)}$ of areas are random variables, as we have seen above.
- **Monte Carlo variability:** the sequence of live points and of the discarded likelihood values has also a random nature.

These sources of uncertainty are intrinsic to the standard ideal NS procedure. Under the ideal assumption of exact and independent sampling from the constrained priors, in the literature, a classical approximation for the uncertainty in the logarithm of the evidence is

$$\text{std} \left[\log \hat{Z}_{\text{ev}} \right] \approx \sqrt{\frac{H}{N}}, \quad (44)$$

where N is the number of live points and we have denoted as

$$H = \int p(\boldsymbol{\theta}|\mathbf{y}) \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{g(\boldsymbol{\theta})} d\boldsymbol{\theta} = \int \bar{\pi}(\boldsymbol{\theta}) \log \frac{\bar{\pi}(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} d\boldsymbol{\theta} = \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})} [\log \ell(\mathbf{y}|\boldsymbol{\theta})] - \log Z_{\text{ev}}, \quad (45)$$

is the Kullback-Leibler (KL) divergence (a.k.a., relative entropy) of the posterior with respect to the prior [11, 25]. The value H measures how much the posterior differs from the prior (in units of *nats* when natural logarithms are used). If the posterior is similar to the prior, H is relatively small. If the posterior is concentrated in a small region of the prior, H is generally larger. Hence, the quantity e^{-H} can be interpreted as an approximate effective fraction of prior mass associated with the posterior concentration. In other words, it is the amount of information lost when $g(\boldsymbol{\theta})$ is used to approximate $\bar{\pi}(\boldsymbol{\theta})$. Hence, the factors $\text{Var} \left[\log \hat{Z}_{\text{ev}} \right] \propto \frac{1}{N}$, and $\text{std} \left[\log \hat{Z}_{\text{ev}} \right] \propto N^{-1/2}$ that are the familiar Monte Carlo (MC) rates.

In practice, the new live point is often generated using MCMC methods (e.g., slice sampling). A NS procedure based on MCMC replacement may fail to explore the full constrained region if it mixes poorly, remains trapped in one mode, or uses too few transitions. In that case, the new added point may not be distributed according to the required constrained prior. This

can introduce *systematic error or bias* in the evidence estimate, in addition to the stochastic uncertainty described above. Repeated independent runs can help assess stochastic variability, while the adequacy of constrained-prior sampling requires separate and more specific diagnostic checks. Consequently, uncertainty estimates based on ideal NS theory should be interpreted cautiously.

5.6 NS behavior varying N

Increasing the number of live points N generally improves the statistical accuracy of NS but, clearly, increases the computational cost. Indeed, recall the theoretical result above $\text{std} \left[\log a_{\max}^{(i)} \right] = \frac{\sqrt{i}}{N}$: for a fixed iteration i , the uncertainty in log prior mass decreases as $1/N$. However, given a suitable stopping rule, larger N also changes the number of iterations required to reach a given prior mass (i.e., given a stopping rule, varying N changes the total number of iterations I). Given the expression $\text{std} \left[\log a_{\max}^{(i)} \right] = \frac{\sqrt{i}}{N}$ and a fixed (normalized) area value $a \in (0, 1]$, the expected number of iterations needed to reach the value a is approximately:

$$i^* \approx -N \lceil \log a \rceil,$$

where $\lceil a \rceil$ represents the smallest integer that is greater than or equal to a . The standard deviation of $\log a_{\max}^{(i)}$ at this iteration i^* is then approximately

$$\text{std} \left[\log a_{\max}^{(i^*)} \right] \approx \frac{\sqrt{-N \lceil \log a \rceil}}{N} = \sqrt{\frac{-\lceil \log a \rceil}{N}}.$$

Thus, at a comparable compression level, the uncertainty in terms of std decreases again at the familiar Monte Carlo rate $N^{-1/2}$, rather than N^{-1} [11, 25]. As remarked previously, even in the idealized NS procedure, the computational cost also increases as N grows. Indeed, even under exact/ideal sampling of the constrained priors, the expected log prior mass after i iterations is

approximately $\mathbb{E} \left[\log a_{\max}^{(i)} \right] = -\frac{i}{N}$. The posterior density typically assigns substantial mass to regions whose log-prior mass is of the order of $-H$, where H is the KL divergence given in Eq. 45 [11, 23, 25]. This motivates the rough estimate

$$\hat{I} \approx N[H],$$

for the number of iterations needed to reach the region contributing most of the evidence. Moreover, when ideal sampling from the constrained prior distributions is not feasible, the computational cost can increase further, particularly when methods such as rejection sampling are employed. On the other hand, increasing the number of live points N improves the initial coverage of the parameter space and, consequently, the probability of capturing distinct posterior modes. This can be particularly important for multimodal posteriors. Indeed, a mode with very small prior mass may still contribute substantially to the evidence, and hence have a large posterior mass, if the likelihood within that mode is sufficiently high.

Therefore, increasing N generally improves statistical accuracy, but also increases the number of iterations required to achieve a pre-established level of prior-mass compression.³ Moreover, a larger number of live points improves the initial coverage of low-prior-mass modes, although it does not guarantee their subsequent exploration. These theoretical considerations must be distinguished from implementation-specific errors due to imperfect constrained-prior sampling [23, 25].

³Prior-mass compression refers to the progressive reduction of the remaining prior mass as the likelihood threshold increases, hence restricting the exploration to increasingly smaller regions of the parameter space with higher likelihood values.

6 Relationship with importance sampling (IS)

6.1 NS and IS estimators of the evidence as weighted likelihood values

We can rewrite Eq. (26) as

$$\boxed{\begin{aligned}\widehat{Z}_{\text{ev}} &= \sum_{i=1}^I \underbrace{(a_{\text{max}}^{(i-1)} - a_{\text{max}}^{(i)})}_{\gamma_i} \underbrace{\lambda_{\text{min}}^{(i)}}_{\ell(\mathbf{y}|\boldsymbol{\theta}_{\text{rem}}^{(i)})}, \\ \widehat{Z}_{\text{ev}} &= \sum_{i=1}^I \gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\text{rem}}^{(i)}),\end{aligned}} \quad (46)$$

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

$$\begin{aligned}\sum_{i=1}^I \gamma_i &= (a_{\text{max}}^{(0)} - a_{\text{max}}^{(1)}) + (a_{\text{max}}^{(1)} - a_{\text{max}}^{(2)}) + (a_{\text{max}}^{(2)} - a_{\text{max}}^{(3)}) + \dots + (a_{\text{max}}^{(I-1)} - a_{\text{max}}^{(I)}), \\ &= a_{\text{max}}^{(0)} - a_{\text{max}}^{(I)}, \\ &= 1 - a_{\text{max}}^{(I)}.\end{aligned} \quad (47)$$

if $I \rightarrow \infty$, we have $\sum_{i=1}^I \gamma_i \approx 1$. where we have used $\lim_{I \rightarrow \infty} a_{\text{max}}^{(I)} = a_{\text{max}}^{(\infty)} = 0$. From the expression above, it is apparent that $\widehat{Z}$ is a linear (asymptotically convex, for $I \rightarrow \infty$) combination of weights γ_i . In [32], the authors describe an importance sampling (IS) estimator with a similar form for the marginal likelihood Z_{ev} :

$$\boxed{\widehat{Z}_{\text{ev}} = \sum_{i=1}^I \rho_i \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad \{\boldsymbol{\theta}_i\}_{i=1}^I \sim q(\boldsymbol{\theta})}, \quad (48)$$

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

Remark 15. We have already pointed out that NS is an automatic quadrature scheme since $a_{max}^{(i)}$, $\lambda_{min}^{(i)}$ are chosen by the NS procedure, as stated in Remark 8: the ordinates $\lambda_{min}^{(i)}$ are selected by a stochastic procedure, and the values $a_{max}^{(i)}$ are estimated by the deterministic formula $\hat{a}_i = \exp\left(-\frac{i}{N}\right)$. Hence, $\gamma_i \approx (e^{-\frac{i-1}{N}} - e^{-\frac{i}{N}})$ are also deterministic weights, as the IS weights that are deterministic $\rho_i = \frac{g(\theta_i)}{I_q(\theta_i)}$, conditioned to θ_i .

6.2 NS estimators of posterior expectations

The NS procedure can be also employed for estimating generic posterior expectations, i.e., $M = \mathbb{E}_{\bar{\pi}}[f(\theta)] = \int_{\Theta} f(\theta) \bar{\pi}(\theta) d\theta$ where $\bar{\pi}(\theta) = p(\theta|\mathbf{y}) = \frac{1}{Z_{ev}} \ell(\mathbf{y}|\theta) g(\theta)$.

Remark 16. Note that the product NS weights $\gamma_i = \Delta a_{max} = a_{max}^{(i-1)} - a_{max}^{(i)}$ and $\lambda_{min}^{(i)} = \ell(\mathbf{y}|\theta_{rem}^{(i)})$, i.e., $\gamma_i \lambda_{min}^{(i)}$, represent a portion of posterior mass: more precisely, a quadrature contribution associated with an interval of prior mass.

Hence, although we have not a partition of θ -space (see Remark 13), we can assert that the product $\gamma_i \ell(\mathbf{y}|\theta_{rem}^{(i)})$ represents a portion of area below the unnormalized posterior $\pi(\theta)$. Indeed, the sum of this products provides an approximation of $Z_{ev} = \int_{\Theta} \pi(\theta) d\theta$, i.e., $\hat{Z}_{ev} = \sum_{i=1}^I \gamma_i \ell(\mathbf{y}|\theta_{rem}^{(i)})$. Thus, the normalized weights

$$\bar{w}_i = \frac{\gamma_i \ell(\mathbf{y}|\theta_{rem}^{(i)})}{\hat{Z}_{ev}} = \frac{\gamma_i \ell(\mathbf{y}|\theta_{rem}^{(i)})}{\sum_{k=1}^I \gamma_k \ell(\mathbf{y}|\theta_{rem}^{(k)})}, \quad (49)$$

allow the approximation of the posterior measure as

$$\hat{\pi}(\theta) = \sum_{i=1}^I \bar{w}_i \delta(\theta - \theta_{rem}^{(i)}). \quad (50)$$

where $\sum_{i=1}^I \bar{w}_i = 1$. Hence, as an IS scheme, the NS method can be also applied to approximate generic integral involving the posterior distribution (as a generalized moment),

$$M = \mathbb{E}_{\bar{\pi}}[f(\boldsymbol{\theta})] = \int_{\Theta} f(\boldsymbol{\theta}) \bar{\pi}(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \widehat{M}_{\text{NS}} = \frac{1}{\widehat{Z}_{\text{ev}}} \sum_{i=1}^I \gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\text{rem}}^{(i)}) f(\boldsymbol{\theta}_{\text{rem}}^{(i)}) = \sum_{i=1}^I \bar{w}_i f(\boldsymbol{\theta}_{\text{rem}}^{(i)}),$$

where $f(\boldsymbol{\theta})$ is any integrable function w.r.t. $\bar{\pi}(\boldsymbol{\theta})$. If the constrained-prior sampling is exact, the estimator above is consistent for $I \rightarrow \infty$. A more refined version for a finite I value, we should include the contribution of the last population of live points as in Eq. (43). Following Remark 14, we can write:

$$\widehat{M}_{\text{NS}} = \frac{\sum_{i=1}^I \gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\text{rem}}^{(i)}) f(\boldsymbol{\theta}_{\text{rem}}^{(i)}) + \frac{a_{\text{max}}^{(I)}}{N} \sum_{n=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)}) f(\boldsymbol{\theta}_n^{(I)})}{\sum_{i=1}^I \gamma_i \ell(\mathbf{y}|\boldsymbol{\theta}_{\text{rem}}^{(i)}) + \frac{a_{\text{max}}^{(I)}}{N} \sum_{n=1}^N \ell(\mathbf{y}|\boldsymbol{\theta}_n^{(I)})}. \quad (51)$$

The numerator estimates the integral with respect to f with respect to the unnormalized posterior $\pi(\boldsymbol{\theta})$, while the denominator estimates the evidence. Including the final live-point terms in both ensures that the same finite-run approximation is used for the numerator and denominator.

6.3 Equivalent proposal density: importance nested sampling [6, 14]

Let us compare Eqs. (46) - (48). The NS estimator can be interpreted as a IS estimator of the form (48), using a “sophisticated” proposal density such that $\rho_i = \gamma_i = a_{\text{max}}^{(i-1)} - a_{\text{max}}^{(i)}$. Generally, 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 computed by the formula $\gamma_i = a_{\text{max}}^{(i-1)} - a_{\text{max}}^{(i)}$. The authors in [14] one or more ellipsoidal distributions are used to construct an approximation of the likelihood contours defined by the inequalities $\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}$. Then, in [6, 14], the authors introduce a *pseudo-proposal density* $q^*(\boldsymbol{\theta})$ that is an equivalent proposal density of all generated points in a NS procedure, considering the ellipsoidal approximations of the likelihood contours. Hence, the the evidence integral could be approximated using the analytic form of an IS estimator

using $q^*(\boldsymbol{\theta})$ as proposal. This equivalent proposal $q^*(\boldsymbol{\theta})$ can be written as a mixture where each component involves the evaluation of an indicator function, that is 1 if $\boldsymbol{\theta}$ lies within the corresponding ellipsoidal approximation or 0 otherwise [6, 14]. A brief and good introduction to this topic is also provided in [49].

6.4 NS-driven importance sampling

Similarly to the idea introduced in [14] with the importance nested sampling, a possible IS scheme inspired by the NS approach is the following. Let us consider $N \cdot I$ samples,

$$\mathcal{P}_i = \{\boldsymbol{\theta}_n^{(i)}\}_{n=1}^N, \quad \boldsymbol{\theta}_n^{(i)} \sim q_i(\boldsymbol{\theta}|\lambda_i) = \beta g(\boldsymbol{\theta}|\lambda_i) + (1 - \beta)q_0(\boldsymbol{\theta}), \quad \beta \in (0, 1],$$

where $g(\boldsymbol{\theta}|\lambda_i)$ are the constrained priors, and $q_0(\boldsymbol{\theta})$ is a generic proposal density $\boldsymbol{\theta} \in \Theta$. Thus, we can assign the following weights to the samples using the deterministic approach [12],

$$w_n^{(i)} = \frac{\pi(\boldsymbol{\theta}_n^{(i)})}{\frac{1}{I} \sum_{i=1}^I q_i(\boldsymbol{\theta}_n^{(i)}|\lambda_i)}. \quad (52)$$

Normalizing them $\bar{w}_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^I \sum_{k=1}^N w_k^{(j)}}$, we can employ these normalized weights for creating IS estimators of any posterior expectations [12, 31]. The values of λ_i can be selected according to the NS procedure. We therefore refer to this scheme as *NS-driven importance sampling (NS-IS)*, an adaptive IS method in which the adaptation mechanism is driven by NS. To the best of our knowledge, this approach has not previously been proposed in the literature. Moreover, a recent work has reformulated nested sampling within the sequential Monte Carlo framework, establishing consistency results for posterior and evidence estimation with MCMC-generated samples, and unbiasedness of the evidence estimator under suitable assumptions [44].

6.5 The vertical-likelihood perspective given in [41]

The NS approach shares some similarities with other Monte Carlo methods that exploit *vertical*-ordinate information in their sampling procedures, such as *slice sampling* and the *Swendsen-Wang algorithm* [29, 38]. An interesting common description is given [41]. Let us consider the following latent-variable representation of the likelihood $\ell(\mathbf{y}|\boldsymbol{\theta})$, used in several well-known methods in literature:

$$\ell(\mathbf{y}|\boldsymbol{\theta}) = \int_0^\infty \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\} d\lambda = \int_0^{\ell(\mathbf{y}|\boldsymbol{\theta})} d\lambda. \quad (53)$$

where $\mathbb{I}$ is the indicator function defined in Section 3.1. Similar representations of the target density are employed in the fundamental theorem of simulation and the inverse-of-density method [35, Chapter 2]. We can define an extended posterior $\bar{\pi}(\boldsymbol{\theta}, \lambda)$ on the augmented space $\Theta \times \mathbb{R}^+$, such that the original posterior, $\bar{\pi}(\boldsymbol{\theta}) = g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})/Z_{\text{ev}}$, is recovered as marginal density with respect to λ . Specifically, we consider the following joint distribution:

$$\bar{\pi}(\boldsymbol{\theta}, \lambda) = \frac{\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\} g(\boldsymbol{\theta})}{Z_{\text{ev}}}, \quad (54)$$

where one marginal distribution is exactly $\bar{\pi}(\boldsymbol{\theta})$ and the other marginal is

$$\bar{\pi}(\lambda) = \int_{\Theta} \bar{\pi}(\boldsymbol{\theta}, \lambda) d\boldsymbol{\theta} = \frac{1}{Z_{\text{ev}}} \int_{\Theta} \mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\} g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{Z(\lambda)}{Z_{\text{ev}}}, \quad \lambda > 0, \quad (55)$$

where $Z(\lambda) = \int_{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda} g(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the same survival function defined in Eq. (7).

Remark 17. *This is the vertical-likelihood representation connected to nested sampling: the marginal density of the auxiliary variable, $\bar{\pi}(\lambda)$, is proportional to the prior mass above the likelihood level λ .*

The conditional densities are:

$$p(\lambda|\boldsymbol{\theta}) = \frac{\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}}{\ell(\mathbf{y}|\boldsymbol{\theta})}, \quad p(\boldsymbol{\theta}|\lambda) = \frac{\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}g(\boldsymbol{\theta})}{Z(\lambda)}. \quad (56)$$

Remark 18. Note that $p(\boldsymbol{\theta}|\lambda) = g(\boldsymbol{\theta}|\lambda)$ is exactly the constrained prior distribution, restricted in $\{\boldsymbol{\theta} \in \Theta : \ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}$, used in nested sampling. Moreover, note that $p(\lambda|\boldsymbol{\theta})$ represents the density of uniform distribution $\mathcal{U}([0, \ell(\mathbf{y}|\boldsymbol{\theta})])$.

Let now define a weight function $w : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, and let $W(\lambda) = \int_0^\lambda w(s)ds$ be the corresponding cumulative function. We can also define a modified, weighted extended posterior as:

$$\bar{\pi}_w(\boldsymbol{\theta}, \lambda) = \frac{w(\lambda)\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}g(\boldsymbol{\theta})}{Z_w}, \quad (57)$$

where

$$\begin{aligned} Z_w &= \int_{\Theta} \int_0^\infty w(\lambda)\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}g(\boldsymbol{\theta})d\boldsymbol{\theta}d\lambda, \\ &= \int_{\Theta} \underbrace{\left[\int_0^\infty w(\lambda)\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}d\lambda \right]}_{\Omega_w(\boldsymbol{\theta})} g(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\Theta} \Omega_w(\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}, \end{aligned} \quad (58)$$

and

$$\begin{aligned} \Omega_w(\boldsymbol{\theta}) &= \int_0^\infty w(\lambda)\mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\boldsymbol{\theta})\}d\lambda, \\ &= \int_0^{\ell(\mathbf{y}|\boldsymbol{\theta})} w(\lambda)d\lambda = W(\ell(\mathbf{y}|\boldsymbol{\theta})) = W \circ \ell(\mathbf{y}|\boldsymbol{\theta}), \end{aligned} \quad (59)$$

hence $Z_w = \int_{\Theta} W(\ell(\mathbf{y}|\boldsymbol{\theta}))g(\boldsymbol{\theta})d\boldsymbol{\theta}$. Moreover, with similar derivations we can find the two marginal distributions from the joint $\bar{\pi}_w(\boldsymbol{\theta}, \lambda)$,

$$\bar{\pi}_w(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta}) W(\ell(\mathbf{y}|\boldsymbol{\theta}))}{Z_w}, \quad \bar{\pi}_w(\lambda) = \frac{w(\lambda)Z(\lambda)}{Z_w}, \quad (60)$$

and the conditional pdf are in this case:

$$p_w(\lambda|\boldsymbol{\theta}) = \frac{w(\lambda)\mathbb{I}\{0 \leq \lambda \leq \ell(\mathbf{y}|\boldsymbol{\theta})\}}{W(\ell(\mathbf{y}|\boldsymbol{\theta}))}, \quad p_w(\boldsymbol{\theta}|\lambda) = \frac{g(\boldsymbol{\theta})\mathbb{I}\{\ell(\mathbf{y}|\boldsymbol{\theta}) > \lambda\}}{Z(\lambda)},$$

so that $p_w(\boldsymbol{\theta}|\lambda) = p(\boldsymbol{\theta}|\lambda) = g(\boldsymbol{\theta}|\lambda)$ is again the constrained prior distribution, used in nested sampling.

Vertical importance sampling [41]. Let assume the ideal case that we are able to draw samples $(\boldsymbol{\theta}_n, \lambda_n)$ from $\bar{\pi}_w(\boldsymbol{\theta}, \lambda)$. In this scenario, each component is distributed as the corresponding marginals, i.e., $\boldsymbol{\theta}_n \sim \bar{\pi}_w(\boldsymbol{\theta})$ and $\lambda_n \sim \bar{\pi}_w(\lambda)$. The idea in [41] is to use the marginal pdf $\bar{\pi}_w(\boldsymbol{\theta})$ as proposal density $q(\boldsymbol{\theta})$ for estimating Z_{ev} via importance sampling. Thus, in this case,

$$\hat{Z}_{\text{ev}} = \sum_{i=1}^I \rho_i \ell(\mathbf{y}|\boldsymbol{\theta}_i), \quad \{\boldsymbol{\theta}_i\}_{i=1}^I \sim \bar{\pi}_w(\boldsymbol{\theta}),$$

where

$$\rho_i = \frac{g(\boldsymbol{\theta}_i)}{I \bar{\pi}_w(\boldsymbol{\theta}_i)} = \frac{Z_w}{I W(\ell(\mathbf{y}|\boldsymbol{\theta}_i))}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^I \sim \bar{\pi}_w(\boldsymbol{\theta}), \quad (61)$$

where Z_w is also unknown. However, defining the normalized weights:

$$\bar{\rho}_i = \frac{\rho_i}{\sum_{n=1}^N \rho_n} = \frac{\frac{1}{W(\ell(\mathbf{y}|\boldsymbol{\theta}_i))}}{\sum_{n=1}^N \frac{1}{W(\ell(\mathbf{y}|\boldsymbol{\theta}_n))}}, \quad \{\boldsymbol{\theta}_i\}_{i=1}^I \sim \bar{\pi}_w(\boldsymbol{\theta}), \quad (62)$$

we can then use the final estimator $\hat{Z}_{\text{ev}} = \sum_{i=1}^I \bar{\rho}_i \ell(\mathbf{y}|\boldsymbol{\theta}_i)$ [32]. We have solved the problem of the unknown value of Z_w , but recall that drawing from $\bar{\pi}_w(\boldsymbol{\theta})$ is not straightforward. This task would generally require an additional Monte Carlo method, such as an MCMC algorithm.

Remark 19. Removing the constant values, the weights $\rho_i \propto 1/W(\ell(\mathbf{y}|\boldsymbol{\theta}_i))$ are proportional to the inverse of the value $W(\ell(\mathbf{y}|\boldsymbol{\theta}_i))$. Hence with the approach in [41], the choice of proposal $q(\boldsymbol{\theta})$

over the original space Θ is thereby reduced to the one-dimensional weight function $w(\lambda)$, that generated the function $W(\lambda) = \int_0^\lambda w(s)ds$.

Remark 20. This cumulative function $W(\lambda)$ can be employed to deform/change the likelihood values $W(\ell(\mathbf{y}|\boldsymbol{\theta}))$ and change then properties of the final IS weights, $\rho_i \propto 1/W(\ell(\mathbf{y}|\boldsymbol{\theta}_i))$.

Special case $w(\lambda) = 1$. With $w(\lambda) = 1$, we have:

- $\bar{\pi}_w(\boldsymbol{\theta}, \lambda) = \bar{\pi}(\boldsymbol{\theta}, \lambda)$,
- $\bar{\pi}_w(\boldsymbol{\theta}) = \bar{\pi}(\boldsymbol{\theta})$ (since $W(\lambda) = \lambda$ in this case),
- $Z_w = Z_{\text{ev}}$ and $\bar{\pi}_w(\lambda) = \frac{Z(\lambda)}{Z_{\text{ev}}}$,
- $\rho_i \propto 1/\ell(\mathbf{y}|\boldsymbol{\theta}_i)$, with $\boldsymbol{\theta}_i \sim \bar{\pi}_w(\boldsymbol{\theta}) = \bar{\pi}(\boldsymbol{\theta})$ and then the method reduces to the so-called harmonic-mean estimator [39, 42].

The harmonic mean estimator of Z can have infinite variance [42]. This issue arises because samples $\boldsymbol{\theta}_i \sim \bar{\pi}(\boldsymbol{\theta})$ can have arbitrarily small likelihood values, $\ell(\mathbf{y}|\boldsymbol{\theta}_i) \approx 0$, leading to arbitrarily large values of the corresponding reciprocal-likelihood terms, $\rho_i = 1/\ell(\mathbf{y}|\boldsymbol{\theta}_i)$. Clearly, reducing the variability of the likelihood values can mitigate this effect. In light of this consideration, it is natural to seek a decreasing function $w(\lambda)$, obtaining a tempering effect as shown in the second special case below.

Special case $w(\lambda) = \alpha\lambda^{\alpha-1}$ **with** $\alpha \in [0, 1]$. In this scenario, $W(\lambda) = \lambda^\alpha$, and the marginal density, used as proposal, is a tempered version of the posterior:

$$\bar{\pi}_w(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})\ell(\mathbf{y}|\boldsymbol{\theta})^\alpha}{Z_w}, \quad \alpha \in [0, 1], \quad (63)$$

hence the method becomes the method described in [32, Section 4.3.2]. With $\alpha < 1$, the idea is to construct a proposal distribution by manipulating the posterior density so that it will retain

a similar shape, but with higher variance or heavier tails. This approach is employed in several well-known techniques, such as the stepping-stones algorithm, the power-posterior method and path sampling [17, 19],[32, Section 4].

Remark 21. *Similarly to the NS-IS scheme described in Section 6.4, and in light of the considerations above, the standard NS procedure can be interpreted as using the constrained prior densities as adaptive proposals for generating the live points, with λ acting as the adaptation parameter.*

7 Practical and theoretical limitations in standard NS

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

- **Constrained-prior sampling bias.** Arguably, the most critical task in the implementation of nested sampling is drawing samples from the truncated priors $g(\boldsymbol{\theta}|\lambda)$. Note that the theoretical results rely on the assumption that the live points are independent and identically distributed (i.i.d.) samples from the corresponding constrained priors. If this assumption does not hold, the theoretical foundations of NS may no longer be valid, and the resulting constructions and guarantees may fail. For drawing from $g(\boldsymbol{\theta}|\lambda)$, 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, as remarked below [11]. 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.

- **Several likelihood modes.** The problem discussed above becomes even more severe when the likelihood exhibits multiple modes. In this case, the support of the constrained prior $g(\boldsymbol{\theta}|\lambda)$ may consist of disconnected regions, as illustrated in Figure 2(a), making it more difficult to sample from $g(\boldsymbol{\theta}|\lambda)$. In particular, MCMC algorithms may become trapped within a single region and fail to explore the remaining modes, leading to mode loss.
- **Stochastic uncertainty in the quadrature variables.** Moreover, in the derivation of the NS method we have considered different approximations:

- The value $\tilde{a}_{\max}^{(i)}$ 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 $\exp(-\frac{1}{N})$ 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 $\exp(-\frac{i}{N})$ is perhaps the reason of using the approximation $\frac{N}{N+1} \approx \exp(-\frac{1}{N})$.
- A further approximation $\mathbb{E}[a_{\max}^{(i)}] \approx \mathbb{E}[\tilde{a}_{\max}^{(i)}]\mathbb{E}[a_{\max}^{(i-1)}]$ is also implicitly applied in (35) (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.
- Furthermore, the likelihood thresholds $\lambda_{\min}^{(i)}$ are determined by a stochastic process that also depends on the number of live points N . As discussed in Section 5.6, the choice of N determines a trade-off between grid resolution and computational cost.
- Last but not least, if the likelihood has a plateau (constant region) at a value c , several live points may have exactly the same likelihood c . Consequently, the minimum can

be attained by multiple points, and the algorithm must decide how to handle the tied points. The likelihood threshold may remain unchanged over several iterations since several points must be removed successively (according to some criterion to handle ties), but still $\lambda_{\min}^{(i)} = c$ until the last point $\boldsymbol{\theta}'$ with $\ell(\mathbf{y}|\boldsymbol{\theta}') = c$ is removed (hence, we have the same node $\lambda_{\min}^{(i)} = c$ for several indices i in the quadrature formula). Moreover, the standard derivation for finding the areas $\hat{a}_i$ does not automatically apply in this scenario and requires additional care [16].

- **Possible weight degeneracy.** 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) [13, 34, 33]. So NS posterior sampling may be less efficient than well-tuned Monte Carlo for parameter estimation (such as MCMC or IS schemes).

Hence, as a summary, the values of the areas $a_{\max}^{(i)}$ and the corresponding values $\lambda_{\min}^{(i)}$ in the NS estimator (26) are computed approximately (they are analytically intractable/unknown values). Sampling from the likelihood-constrained priors is the crucial point and the main challenge. Recall that other sources of error mentioned in Sections 5.1 and 5.5 are:

- Deterministic quadrature error (see Section 5.1 and Eq. (28)).
- Finite-termination and residual-evidence error, as discussed in Section 5.4.
- Possible implementation-dependent computational error (numerical precision etc.).

Finally, if some of the assumptions stated in Section 3.1 are not satisfied, the theoretical derivations become *even* more challenging, and many of the results must be carefully revisited. For instance, if the likelihood has constant regions, NS can produce inaccurate estimations of the evidence and posterior densities [16].

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 [36]. The combination with importance sampling has been designed in [11].

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 [15], 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. Other ideas for tackling complex geometry supports and multimodality are given in [4]. The slice sampling method and the Hamiltonian Monte Carlo (HMC) algorithm has been proposed to be used within NS [2, 21] 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* (Diff-NS) 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, Diff-NS 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 Diff-NS 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. Diff-NS 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 [20], 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 [23], 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 [48, 49]. Finally, solutions to the presence of likelihood plateaus are provided in [16]. Table 4 summarizes the main advanced NS schemes in the literature. Broadly speaking, note that most advanced NS methods aim to address the challenge of efficiently sampling from the constrained priors, which is clearly the main computational issue of the standard NS idea. Furthermore, the main review and survey papers on NS, including the present work, are listed in Table 5. This comparison also clarifies the scope and contribution of the present work relative to previous reviews.

Table 4: The main advanced NS approaches in the literature.

Primary challenge	Key idea
Exploration of constrained densities	NS with “Demons” [20]: introduce auxiliary demon variables to facilitate sampling within likelihood-constrained regions.
Multimodality and mode separation	See [2] using advanced MCMC schemes, and see MultiNest [15]: use clustering and ellipsoidal bounds to identify and sample separate regions of the constrained prior.
Complex geometry and multimodality	UltraNest [4]: use neighbourhood-based regions and adaptive region enlargement to construct sampling regions that better represent the geometry of the live points.
High dimensionality	PolyChord [21]: use multivariate slice sampling to generate constrained samples without relying on rejection from bounding volumes.
Adaptive allocation of computational effort	Dynamic NS [47]: adaptive number of live points, fixing/given a computational budget.
Exploration across likelihood levels	Diffusive NS [3]: allow particles to move both up and down a hierarchy of likelihood-constrained distributions, promoting exploration and helping particles escape poorly explored regions.
Complex distributions and expensive likelihood evaluations	nessai [48]: use normalizing flows to learn transformations that facilitate proposals within the constrained prior.
likelihood with plateaus	several solutions are presented in [16]

9 Conclusions

The Nested sampling (NS) technique has attracted considerable attention and has been widely applied in the literature. Its widespread application, particularly in cosmology and astronomy, demonstrates the significant impact and practical success of NS. 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. Another strength of NS is that its theoretical foundations automatically provide tools for assessing the algorithm’s reliability and diagnosing potential problems, including criteria for determining when to stop the computation. However, the validity of these tools depends on the underlying assumptions being satisfied.

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 [9, 10, 11], with the quality of this constrained sampling being crucial to the method’s performance. In this context, from our perspective, 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. Moreover, the combinations of nested sampling, importance sampling, and vertical Monte Carlo discussed in this work may provide a foundation for developing novel methods and stimulating further research.

Acknowledgements

A preprint version of this work can be found at <https://arxiv.org/pdf/2606.17916>. No funding was received to support this work.

References

- [1] G. Ashton, N. Bernstein, J. Buchner, X. Chen, G. Csányi, A. Fowlie, F. Feroz, M. Griffiths, W. Handley, M. Habeck, E. Higson, M. Hobson, A. Lasenby, D. Parkinson, L. B. Pártay, 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. UltraNest - a robust, general purpose Bayesian inference engine. *Journal of Open Source Software*, 6(60):3001, 2021.
- [5] J. Buchner. Nested sampling methods. *Statistics Surveys*, 17:169 – 215, 2023.
- [6] E. Cameron and A. Pettitt. Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis. *Statistical Science*, 29(3):397–419, 2014.
- [7] M.-H. Chen. Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica*, 59(1):16–29, 2005.
- [8] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.
- [9] N. Chopin and C. P. Robert. Comments on nested sampling by John Skilling. *Bayesian Statistics*, 8:491–524, 2007.

- [10] 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.
- [11] N. Chopin and C. P. Robert. Properties of nested sampling. *Biometrika*, 97(3):741–755, 2010.
- [12] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized Multiple Importance Sampling. *Statistical Science*, 34(1):129–155, 2019.
- [13] V. Elvira, L. Martino, and C. P. Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2022.
- [14] F. Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463, 2008.
- [15] 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.
- [16] A. Fowlie, W. Handley, and L. Su. Nested sampling with plateaus. *Monthly Notices of the Royal Astronomical Society*, 503(1):1199–1205, 05 2021.
- [17] 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.
- [18] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.
- [19] A. Gelman and X. L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

- [20] M. Habeck. Nested sampling with demons. *AIP Conference Proceedings*, 1641(1):121–129, 01 2015.
- [21] 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.
- [22] E. Higson, W. Handley, and A. Hobson, M. land Lasenby. Sampling errors in nested sampling parameter estimation. *Bayesian Analysis*, 13(3):873–896, 2018.
- [23] 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.
- [24] E. Higson, W. Handley, M. Hobson, and A. Lasenby. nestcheck: diagnostic tests for nested sampling calculations. *Monthly Notices of the Royal Astronomical Society*, 483(2):2044–2056, 02 2019.
- [25] C. R. Keeton. On statistical uncertainty in nested sampling. *Monthly Notices of the Royal Astronomical Society*, 414(2):1418–1426, 06 2011.
- [26] F. Keskin. Nested sampling: A case study in parameter estimation. In *Bayesian Inference - Recent Trends*. IntechOpen, 2023.
- [27] 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.
- [28] Jonas Latz, Doris Schneider, and Philipp Wacker. Nested sampling for uncertainty quantification and rare event estimation. *SIAM Journal on Scientific Computing*, 2024.
- [29] F. Liang, C. Liu, and R. Carroll. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley Series in Computational Statistics, England, 2010.

- [30] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.
- [31] F. Llorente and L. Martino. Optimality in importance sampling: a gentle survey. *arXiv:2502.07396*, pages 1–65, 2026.
- [32] 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.
- [33] L. Martino and V. Elvira. Effective sample size approximations as entropy measures. *Computational Statistics*, 40(9):5433–5464, Dec 2025.
- [34] L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.
- [35] L. Martino, D. Luengo, and J. Míguez. Independent random sampling methods. *Springer*, 2018.
- [36] 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.
- [37] P. Mukherjee, D. Parkinson, and A. R. Liddle. A nested sampling algorithm for cosmological model selection. *The Astrophysical Journal*, 638(2):L51, 2006.
- [38] R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–741, 2003.
- [39] M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.
- [40] L. B. Pártay, G. Csányi, and N. Bernstein. Nested sampling for materials. *The European Physical Journal B*, 94:159, 2021.

- [41] N. G. Polson and J. G. Scott. Vertical-likelihood Monte Carlo. *arXiv preprint arXiv:1409.3601*, 2014.
- [42] A. E. Raftery, M. A. Newton, J. M. Satagopan, and P. N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In J. M. et al. Bernardo, editor, *Bayesian Statistics 8: Proceedings of the Eighth Valencia International Meeting, 2006*. Oxford University Press, 2007.
- [43] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.
- [44] R. Salomone, L. F. South, C. Drovandi, D. P. Kroese, and A. M. Johansen. Unbiased and Consistent Nested Sampling via Sequential Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 87(4):1221–1238, 2025.
- [45] J. Skilling. Nested sampling. *AIP Conference Proceedings*, 735(1):395–405, 11 2004.
- [46] J. Skilling. Nested sampling for general Bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.
- [47] J. S. Speagle. dynesty: A dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, 493(3):3132–3158, 04 2020.
- [48] M. J. Williams, J. Veitch, and C. Messenger. Nested sampling with normalizing flows for gravitational-wave inference. *Physical Review D*, 103(10):103006, 2021.
- [49] M. J. Williams, J. Veitch, and C. Messenger. Importance nested sampling with normalising flows. *Machine Learning: Science and Technology*, 4(3):035011, 2023.
- [50] Z. Zhao and T. A. Severini. Integrated likelihood computation methods. *Computational Statistics*, 32(1):281–313, 2017.

Table 5: Reviews, surveys, and tutorials on NS.

Reference	Year	Scope	Main focus and contribution
[5]	2023	General methodological review	Systematic review of NS algorithms, constrained-prior sampling, parallelization, termination criteria, uncertainty estimation, and diagnostic procedures.
[1]	2022	General tutorial and review	Introduction to NS theory and practical implementation, including high-dimensional sampling algorithms, applications, best practices, and limitations.
[28]	2024	Uncertainty quantification	Studies NS as a probabilistic quadrature method and develops theoretical results relevant to uncertainty quantification and rare-event estimation.
[9, 10]	2007	Commentary and critical discussion	Studies different theoretical aspects and practical applicability of NS, highlighting limitations.
[11]	2010	Theoretical and asymptotic analysis	Studies convergence of the evidence estimator, its error rate, and the effect of parameter dimension; also discusses alternatives to MCMC for generating constrained samples.
This work	2026	Critical guide and theoretical review	Examines the theoretical foundations of NS, with particular emphasis on the complete theoretical derivations, assumptions, and potential limitations. The connections IS and vertical MC are examined in particular detail.
[26]	2023	Application-oriented paper	Illustrates the application of NS to direction-of-arrival estimation, discussing practical implementation and parameter estimation.
[40]	2021	Materials science	Reviews adaptations of NS for sampling potential energy surfaces, calculating partition functions, and estimating thermodynamic properties, including parallel and variable-cell extensions.