Nested Sampling And Likelihood Plateaus

Doris Schittenhelm\textsuperscript{1} and Philipp Wacker\textsuperscript{1}

\textsuperscript{1}Friedrich-Alexander-Universität Erlangen-Nürnberg, Department of Mathematics, Cauerstraße 11, 91058 Erlangen, Germany

May 19, 2020

Abstract

The main idea of nested sampling is to substitute the high-dimensional likelihood integral over the parameter space $\Omega$ by an integral over the unit line $[0, 1]$ by employing a push-forward with respect to a suitable transformation. For this substitution, it is often implicitly or explicitly assumed that samples from the prior are uniformly distributed along this unit line after having been mapped by this transformation. We show that this assumption is wrong, especially in the case of a likelihood function with plateaus. Nevertheless, we show that the substitution enacted by nested sampling works because of more interesting reasons which we lay out. Although this means that analytically, nested sampling can deal with plateaus in the likelihood function, the actual performance of the algorithm suffers under such a setting and the method fails to approximate the evidence, mean and variance appropriately. We suggest a robust implementation of nested sampling by a simple decomposition idea which demonstrably overcomes this issue.

Keywords: nested sampling, evidence, Bayesian statistics, generalized inverses

1 Introduction

Nested sampling was introduced by Skilling in \cite{17} and has been applied in astronomy \cite{13, 7, 6, 14}, biomathematics \cite{11, 4, 15} and other fields with great
success. From a bird’s eye view, nested sampling is a method of computing the evidence $\int_{\Omega} \mathcal{L}(x) \, d\mu(x)$ by doing

$$\int_{\Omega} \mathcal{L}(x) \, d\mu(x) = \int_0^1 \hat{\mathcal{L}}(X) \, dX \approx \text{Monte Carlo approximation.} \quad (1)$$

Here, $\hat{\mathcal{L}}$ is the “overloaded” form of the likelihood $\mathcal{L}$ as defined in [17] because we will explicitly distinguish between $\mathcal{L}$ and $\hat{\mathcal{L}}$.

There is already an extensive body of work regarding the performance of the second step, i.e. the quality of the Monte Carlo approximation of the 1d integral $\int_0^1 \hat{\mathcal{L}}(X) \, dX$ and a series of improvements to Skilling’s original method has emerged, for example in [11, 8, 10, 2]. To the best of our knowledge there has not been much discussion on the validity of the first step, i.e. the substitution of the integral over $\Omega$ with an integral over the line $[0, 1]$. At the heart of the matter this is just calculation of the push-forward measure, but the details are surprisingly involved and contain many steps that “almost” fail to hold.

The main difficulty is “plateaus” in $\mathcal{L}$ (or rather: $\hat{\mathcal{L}}$). This was already hinted at in Skilling’s original publication [17] where he discusses “cliffs” – which he judges at being non-problematic – and “plateaus”. Skilling assesses plateaus as difficult, but rules “[...] even so, it may be possible to generate [new active samples] efficiently.” In other words, Skilling states that the existence of plateaus makes the Monte Carlo method more prone to performance problems which can be overcome by a cleverer method.

We will show that plateaus in $\hat{\mathcal{L}}$ are also fundamentally mathematically problematic: If there is a level $\alpha \in \mathbb{R}$ such that $\mu(\mathcal{L} = \alpha) > 0$, then the following implicit assumption about nested sampling is violated:

“Samples from the prior which are plotted in a $X$-$\mathcal{L}$-diagram are uniformly distributed along the axis $[0, 1]$.” This is what Sivia and Skilling mean when they write “In terms of $\xi$, the objects are uniformly sampled subject to the constraint $\xi < \xi^\star$” [16 section 9.2].

We show that this is wrong in general; but although this seems to invalidate the main paradigm of nested sampling, we prove that the integral substitution which is the first equality in (1) is still correct.

Then we demonstrate that this sadly does not reflect on the actual performance of nested sampling in such a situation as the evidence is computed badly mainly due to the fact that the estimation of the shrinkage ratio of the accessible prior mass is not valid any longer which is based on the violated uniformity assumption. But insights from the proof of the main statement allow us to propose a slightly modified algorithm which improves its
performance to the level of nested sampling for non-problematic (i.e. non-
plateauing likelihoods).

We suspect that the techniques employed (especially lemma\cite{1}) have some
connections with the notion of generalized inverse distribution functions \cite{3, 12, 5} which is slightly (but unintentionally) obfuscated by the fact that
we consider non-increasing instead of non-decreasing functions. We haven’t
pursued this direction but it might be interesting to combine both ideas.

\section{The main paradigm of nested sampling}

Let $(\Omega, \mathcal{A}, \mu)$ be a probability space where we interpret $\mu$ as a prior probability
measure of parameters $x \in \Omega$. Sometimes (but not for our main results) we
will assume that $\mu$ has a Lebesgue density $\pi$. Furthermore, we consider a
likelihood $L : \Omega \rightarrow \mathbb{R}$. This is usually due to the arrival of measurement
data giving implicit information about possible parameter values. Hence,
by Bayes’ Law, the posterior probability is then given by $\mu_p(dx) = Z^{-1} \cdot
L(x) \cdot \mu(dx)$. The normalization constant $Z = \int_{\Omega} L(x) \mu(dx)$ is called the
evidence and is used for example in Bayesian model selection. It is usually a
non-trivial matter to compute $Z$. Nested sampling is a method specifically
designed for doing this.

The basic idea of nested sampling is to swap the integration domain
from $\Omega$ for the interval $[0, 1]$ by constructing a specific transformation $\Phi$ and
conducting basically a change of variables (or rather, a push-forward). Then
the high-dimensional integration over the likelihood (which constitutes the
evidence) becomes an integral of a function over the line $[0, 1]$, which can
more readily be approximated by a Monte-Carlo method. Concretely, nested
sampling evaluates $Z$ by computing the right-hand-side of

$$
\int_{\Omega} L(x) \, d\mu(x) = \int_{0}^{1} \tilde{L}(X) \, dX
$$

instead of the left-hand-side (which is the definition of $Z$). Then this one-
dimensional integral is efficiently approximated by a Monte-Carlo method
(using “active” and “dead” samples and a clever way of estimating probabil-
ities). A more in-depth explanation of this remarkable idea can be found in
for example \cite{17, 16} and won’t concern us here because this is not the focus
of this manuscript.

This is the main idea of nested sampling: Folding (or nesting) all points
on a level set of constant $\mathcal{L}$ into one point on the interval $[0, 1]$ and hereby
reducing dimensional complexity. We will concern ourselves with this trans-
formation of integration and we show specifically under which constraints this
construction works (regardless of the Monte-Carlo approximation procedure used for evaluation of the 1d integral).

We demonstrate next how we need to choose the transformation and the function \( \tilde{L} \) such that identity (2) has a chance to hold. First, we define

\[
X(\lambda) := \mu(\{z \in \Omega : L(z) > \lambda\}),
\]

i.e. the \( \mu \)-measure of the \( \lambda \)-super-level-sets. Then we introduce a mapping \( \Phi : (\Omega, B^d) \to ([0, 1], B([0, 1])) \) with

\[
\Phi(x) = X(L(x))
= \int_{\{z \in \Omega : L(z) > L(x)\}} \pi(z) \, dz
= \int_{\{z \in \Omega : L(z) > L(x)\}} d\mu(z)
= \mu(\{z \in \Omega : L(z) > L(x)\})
\]

which is visualized in Figure 1 for \( \Omega = \mathbb{R} \), a Gaussian measure \( \mu \) and a Gaussian-type likelihood. We remark that \( \Omega = \mathbb{R} \) is chosen just for visualization reasons and all considerations of the proof are also valid in higher dimensions.

In addition to \( \Phi(x) \), we define the pushforward measure \( \Phi_*\mu \) on \([0, 1]\) (see Figure 2 for visualization). This transformation \( \Phi \) allows us to shift the integration in \( \Omega \) space to integration in the image of \( \Phi \), which is the interval \([0, 1]\).

Next we need to specify the form of \( \tilde{L} \), the new function which is integrated over the line \([0, 1]\). Skilling also calls it \( L \) (an “overloaded form”) and identifies it with the inverse function of \( X \circ L \), but things are not as...
Figure 2: Visualization of relation of probability measure $\mu$ on $\Omega$ and the pushforward measure $\Phi_{\#\mu}$ on $[0,1]$. We call the idea of transforming the evidence integral on $\Omega$ into an integral on $[0,1]$ the main paradigm of nested sampling.

straightforward as they seem: In general, $X \circ L$ has no inverse. For now, we just define

$$\hat{L}(\xi) = \sup\{\lambda \in \text{im } L : X(\lambda) > \xi\}.$$  

$\hat{L}$ is a generalized inverse for $X$. In Section 4 we will show that

$$\hat{L}(X(L(x))) = L(x) \text{ } \mu \text{-a.s.}$$

holds. In the general case we cannot hope for this identity to hold everywhere (i.e. for all $x \in \Omega$) and then of course $\hat{L}(X(\lambda)) \neq \lambda$ for general $\lambda \in \mathbb{R}$.

Now we are able to derive the integral transformation which lies at the heart of nested sampling.

$$\int_{\Omega} L(x) \, d\mu(x) \overset{\text{sec. 4}}{=} \int_{\Omega} \hat{L}(X(L(x))) \, d\mu(x)$$

$$\overset{\text{def. of } \Phi}{=} \int_{\Omega} (\hat{L} \circ \Phi)(x) \, d\mu(x)$$

$$\overset{\text{def. of } \Phi_{\#\mu}}{=} \int_{[0,1]} \hat{L}(r) \, d(\Phi_{\#\mu})(r)$$

$$\overset{\text{sec. 3}}{=} \int_{0}^{1} \hat{L}(X) \, dX$$

Equation (3) holds because $\hat{L}(X(L(x))) = L(x)$ is true for $\mu$-almost-all $x \in \Omega$ (but not for all $x \in \Omega$). This is discussed in section 4. The equality in (4) is true, as detailed in section 3, but not for obvious reasons: We will show that in general, $\Phi_{\#\mu} \neq \text{Unif}[0,1]$ (which would in particular imply the equality).

All in all, this manuscript mainly proves that these two equalities are true even under adverse circumstances.

5
This can be likened to the following situation: A student tries to calculate the solution of $x^2 = 25$. They “cancel the 2”, i.e. $x^2 = 25 \Rightarrow x = 5$. This result is correct, although the method is not (because the student is not using the proper method, which is using the square root). We will attempt something similar to taking the square root in the context of nested sampling: We will show that nested sampling is well-defined although things are not as easy as they seem.

The last step (4) can be interpreted in another way: $\Phi$ is usually assumed \cite{17, 16} to have the following desirable property: If we generate i.i.d. samples $x_i \in \Omega$ from $\mu$ and map them via $\Phi$ into $[0, 1]$, then the set $\{\Phi(x_i)\}$ consists of samples from a uniform distribution on the set $[0, 1]$ (see also Figure 3). This is equivalent to saying that the measure $\Phi\#\mu$ is a uniform measure on $[0, 1]$, or that the cumulative distribution $\Phi\#\mu([0, r]) = r$. This, in turn, means that the Lebesgue-density of $\Phi\#\mu$ is given by $d\Phi\#\mu = dr$. This means that we can also interpret step (4) to hold if and only if $\Phi$ is a uniform measure. This is what Sivia and Skilling mean when they write “In terms of $\xi$, the objects are uniformly sampled subject to the constraint $\xi < \xi^*$” \cite[section 9.2]{16}. One of the main point of this manuscript is that this is sometimes wrong but that the integral transformation is still valid.

In one paragraph, the results of this manuscript can be summarized as follows:
If $\mathcal{L}$ has a plateau of positive measure $\mu$, then $\Phi_{\#\mu}$ is not the uniform measure on $[0, 1]$. Also, $\tilde{\mathcal{L}}(X(\mathcal{L}(x))) \neq \mathcal{L}(x)$ in general. Still, (3) and (4) are true and thus the main paradigm of nested sampling (2) holds.

**Theorem 1.** Let $(\Omega, \mathcal{B}^d, \mu)$ be a probability space with $\Omega \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$ and $\mathcal{B}^d$ the Borel $\sigma$-algebra on $\Omega$. Let further $\mathcal{L} : (\Omega, \mathcal{B}^d) \to (\mathbb{R}, \mathcal{B})$ be a measurable function bounded from above and below, whereby $\mathcal{B}$ is the Borel $\sigma$-algebra on $\mathbb{R}$. Then

$$\int_{\Omega} \mathcal{L}(x) \, d\mu(x) = \int_{0}^{1} \tilde{\mathcal{L}}(X) \, dX. \tag{5}$$

**Proof.** Follows from (4) with corollary 1.

A brief overview of the remainder of the paper:

- Section 3 explores the form of the push-forward measure $\Phi_{\#\mu}$ in detail. This directly relates to the existence and form of plateaus in the likelihood $\mathcal{L}$.

- Section 4 shows that although (3) is non-trivial (i.e. not valid pointwise inside the integral), it is still true.

- Section 5 demonstrates the failure of the basic version of nested sampling to compute the evidence in a situation where there is a plateau in $\mathcal{L}$. This is to be contrasted with the results of the preceding sections: *Theoretically*, the paradigm nested sampling is well-defined even with plateaus in $\mathcal{L}$, but *practically*, the algorithm struggles.

- Section 6 derives a slight adaptation of the original nested sampling algorithm that can handle plateaus in $\mathcal{L}$. This does not touch the “paradigm” of nested sampling but only its numerical approximation (i.e. the second equality in (4), not the first).

### 3 The form of $\Phi_{\#\mu}$

In this section we will talk about the equality of

$$\int_{0}^{1} \tilde{\mathcal{L}}(r) \, d(\Phi_{\#\mu})(r) = \int_{0}^{1} \tilde{\mathcal{L}}(X) \, dX. \tag{6}$$

7
In the simplest case, this would be true if $\Phi \# \mu$ is the uniform measure on $[0,1]$, because then $\frac{d\Phi \# \mu}{dr} = 1$. This is equivalent to saying that i.i.d $\mu$-samples $\{x_i\}_i \in \Omega$ become uniform samples $\{\Phi(x_i)\}_i$ in $[0,1]$ via the mapping $\Phi$.

We will show that $\Phi \# \mu = \text{Unif}[0,1]$ if and only if there are no plateaus of $L$ which have positive $\mu$-measure (i.e. $L \# \mu$ has not atoms). In the case where the push-forward $\Phi \# \mu$ is not the uniform measure, we will show that equality (6) is still true.

The cumulative distribution function of $\Phi \# \mu$ is given by

$$
\Phi \# \mu([0,\alpha]) = \mu(\{x \in \Omega : \Phi(x) \in [0,\alpha]\})
$$

Equation (7)

$$
= \mu(\{x \in \Omega : X(L(x)) \in [0,\alpha]\})
$$

Equality (8)

$$
= \mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\})
$$

If

$$
\mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\}) = \alpha,
$$

then and only then $\Phi \# \mu = \text{Unif}[0,1]$.

In this section we derive conditions for the validity of Equation (8). Key statement is the following lemma.

**Lemma 1.** Let $\mu$ be a measure on a measurable space $(\Omega, \mathcal{A})$, $\Omega \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, and $L : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ a measurable mapping which is bounded from above and below, i.e. $-\infty < \inf L \leq \sup L < \infty$. We consider the push-forward of $\mu$ via $L$, denoted by $L \# \mu$ (which now is a measure on $(\mathbb{R}, \mathcal{B})$). We assume that we know all positions $r_i \in \mathbb{R}$ at which the discrete part of $L \# \mu$ has a non-vanishing contribution, in particular we assume that we know the $(r_i)_{i=1}^N$, $(\alpha_i)_{i=1}^N$, and $(\Delta_i)_{i=1}^N$ such that

$$
L \# \mu(\{r_i\}) = \mu(\{z \in \Omega : L(z) = r_i\}) = \Delta_i
$$

$$
L \# \mu((r_i, \infty)) = \mu(\{z \in \Omega : L(z) > r_i\}) = \alpha_i.
$$

We assume that there are only finitely many (i.e. $N$) of such atoms and that $L \# \mu$ is non-singular everywhere else. Then for any $\alpha \in [0,1]$,

$$
\mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\}) = \begin{cases} 
\alpha_i + \Delta_i & \text{for } \alpha \in [\alpha_i, \alpha_i + \Delta_i)
\end{cases}
$$

$$
\alpha & \text{else.}
$$

For a visualization of this and the quantities $\alpha_i, \Delta_i$, see Figure 4.
Figure 4: Row-wise examples of measure $\mu$ and likelihood function in the context of Lemma 1. Apparently, $\mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\})$ is not uniform e.g. if $L$ has plateaus on the support of $\mu$, neither in case 0 (row 1) nor in case AB (row 2).
Proof. We sketch the proof formally first: Key ingredients will be two functions which, concatenated, yield the result. First, we will define \( \psi(r) = \mu(\{z \in \Omega : L(z) \geq r\} \) and then we define \( \alpha \mapsto r_\alpha := \inf\{r \in \mathbb{R} : \psi(r) \leq \alpha\} \) (a kind of generalized inverse for \( \psi \)). Note that \( \psi \) is similar to the function \( X \) defined in section 2, but with \( \geq \) instead of \( > \), which we will need. Furthermore, \( r_\alpha = \hat{L}(\alpha) \), although this is not obvious at this point and will be proven in step 2.7. We will be able to prove that

\[
\psi(r_\alpha) = \begin{cases} 
\alpha_i + \Delta_i & \text{for } \alpha \in [\alpha_i, \alpha_i + \Delta_i) \\
\alpha & \text{else.}
\end{cases}
\]

and also that \( \mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\}) = \mu(\{x \in \Omega : L(x) \geq r_\alpha\}) = \psi(r_\alpha) \), which, in combination, yield the result.

Due to a few technicalities, there are some pathological cases that need to be taken care of individually. In order to increase readability, we suggest to skip passages marked grey on first reading because they only contain boundary cases or technicalities. These are necessary but are bound to distract from the guiding thread.

We show the result in a series of steps. First, we will need to fix notation for the actual proof.

We start by ordering \( r_1 < r_2 < \cdots \). Note that two \( r_i, r_j, i \neq j \) cannot be identical or we could just drop one of them. With this ordering, we obtain \( \alpha_1 + \Delta_1 > \alpha_1 \geq \alpha_2 + \Delta_2 > \alpha_2 \geq \cdots \), i.e.

\[
\alpha_i + \Delta_i > \alpha_i \\
\alpha_i \geq \alpha_{i+1} + \Delta_{i+1}
\]

Indeed, \( \alpha_k + \Delta_k > \alpha_k \) because \( \Delta_k > 0 \) and \( \alpha_k = \mathcal{L}#\mu([r_k, \infty)) \geq \mathcal{L}#\mu([r_{k+1}, \infty)) = \alpha_{k+1} + \Delta_{k+1} \) because \( r_{k+1} > r_k \). Here, it is possible to have equality, i.e. a situation where \( \mathcal{L}#\mu((r_k, \infty)) = \mathcal{L}#\mu([r_{k+1}, \infty)) \) or equivalently \( \alpha_k = \alpha_{k+1} + \Delta_{k+1} \).

From this point on, we will without loss of generality always consider this ordering of the \( r_i \) and \( \alpha_i \).

There are two special cases to consider depending on whether the infimum and the supremum of \( L \) constitute plateaus with positive measure themselves: Special Case A will be when \( r_1 = \inf L \) which holds if and only if \( \alpha_1 + \Delta_1 = 1 \). Special Case B is when \( r_N = \sup L \) (i.e. if and only if \( \alpha_N = 0 \)). Any of those two cases can be present (similar to blood types), i.e. we will always have four cases to keep in mind: A, B, AB, and 0 (i.e. neither A nor B).
Before we can continue, we need to define specific decompositions for the intervals \([\inf L, \sup L]\) and \([0, 1]\) (the range of possible probabilities).

We set
\[
I_i = (r_{i}, r_{i+1}], \quad i = 1, \ldots, N - 1.
\]

The leftmost and rightmost intervals need to be chosen according to the special case \(\in (A, B, AB, 0)\) we are in:
\[
I_0 = \begin{cases} 
\{\inf L\}, & \text{if } r_1 = \inf L, \text{ i.e. in cases } A, AB \\
[\inf L, r_1], & \text{if } r_1 > \inf L, \text{ i.e. in cases } 0, B
\end{cases}
\]

In the same way, we need to be careful with the other boundary. If \(r_N = \sup L\), then \(I_{N-1} = (r_{N-1}, r_N]\) is the last interval. If \(r_N < \sup L\), we need to additionally define \(I_N := (r_N, \sup L]\). In order to simplify notation, we write
\[
I_N = \begin{cases} 
\emptyset, & \text{if } r_N = \sup L, \text{ i.e. in cases } B, AB \\
(r_N, \sup L], & \text{if } r_N < \sup L, \text{ i.e. in cases } 0, A
\end{cases}
\]

At any rate, \([\inf L, \sup L] = \bigcup_{i=0}^{N} I_i\) is a disjoint decomposition (see Figure 5).

Similarly we define a decomposition of the interval \([0, 1]\): We set
\[
A_i := [\alpha_i, \alpha_i + \Delta_i).
\]

for \(i = 1, \ldots, N\). The “missing parts” are defined via
\[
J_i := \begin{cases} 
[\alpha_{i+1} + \Delta_i, \alpha_i), & \text{if } \alpha_{i+1} + \Delta_i < \alpha_i \\
[\alpha_i), & \text{if } \alpha_{i+1} + \Delta_i = \alpha_i
\end{cases}
\]

for \(i = 1, \ldots, N - 1\).

As before, we need to pay attention to details concerning the “boundary conditions”:
\[
J_0 = \begin{cases} 
\{1\}, & \text{if } r_1 = \inf L, \text{ i.e. in cases } A, AB \\
[\alpha_1 + \Delta_1, 1], & \text{if } r_1 > \inf L, \text{ i.e. in cases } 0, B
\end{cases}
\]

and
\[
J_N = \begin{cases} 
\emptyset, & \text{if } r_N = \sup L, \text{ i.e. in cases } B, AB \\
[0, \alpha_N), & \text{if } r_N < \sup L, \text{ i.e. in cases } 0, A
\end{cases}
\]

A visualization of the intervals is shown in Figure 6.

11
Step 0. The intervals $I_i$ yield a complete and disjoint decomposition\(^{\footnote{We use $\cup$ for union of sets in general and $\bigcup$ explicitly for disjoint unions of sets, i.e. $A \cup B$ contains no points which are in both $A$ and $B$.}}$

$$[\inf \mathcal{L}, \sup \mathcal{L}] = \bigsqcup_{i=0}^{N} I_i.$$  

The intervals $A_i := [\alpha_i, \alpha_i + \Delta_i)$ are disjoint from each other and are “proper” intervals, i.e. $\alpha_i + \Delta_i > \alpha_i$. The $A_i$ and the $J_i$ together form a decomposition:

$$[0, 1] = J_0 \cup \bigsqcup_{i=1}^{N} (A_i \cup J_i).$$

Note that this decomposition is not completely disjoint: $A_i \cap J_i = \{\alpha_i\} = J_i$ if and only if $\alpha_{i+1} + \Delta_{i+1} = \alpha_i$. Apart from that, all sets involved are disjoint from each other, in particular $(A_i \cup J_i) \cap (A_j \cup J_j) = \emptyset$ for $i \neq j$.

Proof of Step 0 Proven in the text above. \hfill \Box

Step 1. We consider first the map $\psi : r \mapsto \mu(\{z \in \Omega : \mathcal{L}(z) \geq r\}) = \mathcal{L} \# \mu([r, \infty))$. This mapping has the following properties:

1.1. $\psi$ is non-increasing.
1.2. $\psi$ is left-continuous at $r = r_i$, with $\lim_{r \searrow r_i} \psi(r) = \psi(r_i) = \alpha_i + \Delta_i$.
1.3. $\psi$ is non-right-continuous at $r = r_i$, because we have $\lim_{r \nearrow r_i} \psi(r) = \alpha_i < \alpha_i + \Delta_i = \psi(r_i)$.
1.4. Any restricted mapping $\psi_i := \psi|_{I_i}$ is a continuous, well-defined, surjective, and non-increasing map $\psi_i : I_i \to J_i$.

(a) case 0

$\begin{align*}
I_0 &= \{\inf \mathcal{L}\} \\
r_1 &= \inf \mathcal{L} \\
I_1 &= \bigr] \left[ \\
r_2 &= \sup \mathcal{L} \\
I_2 &= \bigr] \left[ \\
I_3 &= \emptyset \\
r_3 &= \sup \mathcal{L}
\end{align*}$

(b) case AB

Figure 5: Decomposition of interval $[\inf \mathcal{L}, \sup \mathcal{L}]$ into $I_i$. 

Step 0. The intervals $I_i$ yield a complete and disjoint decomposition

$$[\inf \mathcal{L}, \sup \mathcal{L}] = \bigsqcup_{i=0}^{N} I_i.$$  

The intervals $A_i := [\alpha_i, \alpha_i + \Delta_i)$ are disjoint from each other and are “proper” intervals, i.e. $\alpha_i + \Delta_i > \alpha_i$. The $A_i$ and the $J_i$ together form a decomposition:

$$[0, 1] = J_0 \cup \bigsqcup_{i=1}^{N} (A_i \cup J_i).$$

Note that this decomposition is not completely disjoint: $A_i \cap J_i = \{\alpha_i\} = J_i$ if and only if $\alpha_{i+1} + \Delta_{i+1} = \alpha_i$. Apart from that, all sets involved are disjoint from each other, in particular $(A_i \cup J_i) \cap (A_j \cup J_j) = \emptyset$ for $i \neq j$.

Proof of Step 0 Proven in the text above. \hfill \Box

Step 1. We consider first the map $\psi : r \mapsto \mu(\{z \in \Omega : \mathcal{L}(z) \geq r\}) = \mathcal{L} \# \mu([r, \infty))$. This mapping has the following properties:

1.1. $\psi$ is non-increasing.
1.2. $\psi$ is left-continuous at $r = r_i$, with $\lim_{r \searrow r_i} \psi(r) = \psi(r_i) = \alpha_i + \Delta_i$.
1.3. $\psi$ is non-right-continuous at $r = r_i$, because we have $\lim_{r \nearrow r_i} \psi(r) = \alpha_i < \alpha_i + \Delta_i = \psi(r_i)$.
1.4. Any restricted mapping $\psi_i := \psi|_{I_i}$ is a continuous, well-defined, surjective, and non-increasing map $\psi_i : I_i \to J_i$.

(a) case 0

$\begin{align*}
I_0 &= \{\inf \mathcal{L}\} \\
r_1 &= \inf \mathcal{L} \\
I_1 &= \bigr] \left[ \\
r_2 &= \sup \mathcal{L} \\
I_2 &= \bigr] \left[ \\
I_3 &= \emptyset \\
r_3 &= \sup \mathcal{L}
\end{align*}$

(b) case AB

Figure 5: Decomposition of interval $[\inf \mathcal{L}, \sup \mathcal{L}]$ into $I_i$. 

Step 0. The intervals $I_i$ yield a complete and disjoint decomposition

$$[\inf \mathcal{L}, \sup \mathcal{L}] = \bigsqcup_{i=0}^{N} I_i.$$  

The intervals $A_i := [\alpha_i, \alpha_i + \Delta_i)$ are disjoint from each other and are “proper” intervals, i.e. $\alpha_i + \Delta_i > \alpha_i$. The $A_i$ and the $J_i$ together form a decomposition:

$$[0, 1] = J_0 \cup \bigsqcup_{i=1}^{N} (A_i \cup J_i).$$

Note that this decomposition is not completely disjoint: $A_i \cap J_i = \{\alpha_i\} = J_i$ if and only if $\alpha_{i+1} + \Delta_{i+1} = \alpha_i$. Apart from that, all sets involved are disjoint from each other, in particular $(A_i \cup J_i) \cap (A_j \cup J_j) = \emptyset$ for $i \neq j$.

Proof of Step 0 Proven in the text above. \hfill \Box

Step 1. We consider first the map $\psi : r \mapsto \mu(\{z \in \Omega : \mathcal{L}(z) \geq r\}) = \mathcal{L} \# \mu([r, \infty))$. This mapping has the following properties:

1.1. $\psi$ is non-increasing.
1.2. $\psi$ is left-continuous at $r = r_i$, with $\lim_{r \searrow r_i} \psi(r) = \psi(r_i) = \alpha_i + \Delta_i$.
1.3. $\psi$ is non-right-continuous at $r = r_i$, because we have $\lim_{r \nearrow r_i} \psi(r) = \alpha_i < \alpha_i + \Delta_i = \psi(r_i)$.
1.4. Any restricted mapping $\psi_i := \psi|_{I_i}$ is a continuous, well-defined, surjective, and non-increasing map $\psi_i : I_i \to J_i$.
More precisely we can say the following:

- If $i = 0$, then either the mapping is trivial (in cases A or AB) in the sense that $\psi_0 : \{\inf \mathcal{L}\} \to \{1\}$ or (in cases 0 or B): $\psi_0 : I_0 = [\inf \mathcal{L}, r_1] \to [\alpha_1 + \Delta_1, 1]$ is a continuous, non-increasing and onto mapping with $\psi(r_1) = \alpha_1 + \Delta_1$ and $\psi(\inf \mathcal{L}) = 1$.

- If $i = N$, then either the mapping is trivial (in cases B or AB) in the sense that $\psi_N : \emptyset \to \emptyset$ or (in cases 0, A): $\psi_N : (r_N, \sup \mathcal{L}] \to [0, \alpha_N)$ is a continuous, non-increasing and onto mapping with $\lim_{r \searrow r_N} \psi(r) = \alpha_N$ and $\psi(\sup \mathcal{L}) = 0$.

- If $i = 1, \ldots, N - 1$, then there are two possibilities:
  
  - $\alpha_{i+1} + \Delta_{i+1} = \alpha_i$ and $\psi_i \equiv \alpha_i$, i.e. $\psi_i : I_i \to J_i$ collapses $I_i$ to one point, or
  
  - $\alpha_{i+1} + \Delta_{i+1} < \alpha_i$. Then $\psi_i$ is continuous, non-increasing and surjective as a function $I_i \to J_i$.

  If $i = 1, \ldots, N - 1$, then: $\lim_{r \searrow r_i} \psi(r) = \alpha_i$ and $\psi(r_{i+1}) = \alpha_{i+1} + \Delta_{i+1}$.

These properties can also be seen in Figure 7 and 8 from our example at the beginning (Figure 4).

1.5. (For later purposes): We define a slight variation on $\psi$ given by $\phi : r \mapsto \mu(\{z \in \Omega : \mathcal{L}(z) > r\}) = \mathcal{L}_\# \mu((r, \infty))$. Then $\phi$ is the right-continuous version of $\psi$, i.e. $\phi(r) = \psi(r)$ for any $r \neq r_i$ and $\phi(r_i) = \alpha_i$. Note that $\phi(r) = X(r)$.
(a) Mapping $\psi(r)$. 

$$\psi(r)$$

$1$

$\alpha_1 + \Delta_1$

$\alpha_1 = \alpha_2 + \Delta_2$

$\alpha_2$

$J_2$

$\inf(L)$

$r_1$

$r_2$

$I_2^{\sup(L)}$

$r$

(b) Interval mapping via $\psi(r)$. 

Figure 7: Case 0: $\psi(r)$ and interval mapping from $[\sup L, \inf L]$ to $[0, 1]$ via $\psi(r)$ for the likelihood and the measure shown in Figure 4 (1st row).
(a) Mapping \( \psi(r) \).

(b) Interval mapping via \( \psi(r) \)

Figure 8: Case AB: \( \psi(r) \) and interval mapping from \( [\sup \mathcal{L}, \inf \mathcal{L}] \) to \( [0, 1] \) via \( \psi(r) \) for the likelihood and the measure shown in Figure 4 (2nd row).
Proof of Step [2]  Ad [1.1] Define \( B_r = \{ z \in \Omega : \mathcal{L}(z) \geq r \} \) The monotonicity of \( \psi \) follows from the inclusion of the sets \( B_s \subseteq B_r \) for \( r \leq s \).

Ad [1.2] Let \( (r^{(n)})_{n \in \mathbb{N}} \) be any sequence monotonously increasing with limit \( r_i \) for some \( i \), i.e. \( r^{(n)} \nearrow r_i \). We need to show that \( \lim_{n \to \infty} \psi(r^{(n)}) = \psi(r_i) = \alpha_i + \Delta_i \). Define \( E_n := \{ z \in \Omega : \mathcal{L}(z) \geq r^{(n)} \} \), and thus \( E_n \supseteq E_{n+1} \supseteq \cdots \) and \( \bigcap_n E_n = \{ z \in \Omega : \mathcal{L}(z) \geq r_i \} \). The measure \( \mu \) is continuous from above, which implies the identity marked with \( * \) in the identity \( \lim_{n \to \infty} \psi(r^{(n)}) = \lim_{n \to \infty} \mu(E_n) = \mu(\bigcap_n E_n) = \mu(\{ z \in \Omega : \mathcal{L}(z) \geq r_i \}) \). But this is clear \( \psi \) follows from left-continuity of \( \mu \).

Ad [1.3] Let now \( r^{(n)} \searrow r_i \) monotonously decreasing and define \( F_n := \{ z \in \Omega : \mathcal{L}(z) \geq r^{(n)} \} \). Now \( F_n \subseteq F_{n+1} \subseteq \cdots \) and \( \bigcup_n F_n = \{ z \in \Omega : \mathcal{L}(z) > \inf_n r^{(n)} \} = \{ z \in \Omega : \mathcal{L}(z) > r_i \} \). The measure \( \mu \) is also continuous from below, i.e. \( \lim_{n \to \infty} \psi(r^{(n)}) = \lim \mu(F_n) = \mu(\bigcup_n F_n) = \mu(\{ z \in \Omega : \mathcal{L}(z) > r_i \}) = \alpha_i \) which is strictly less than \( \alpha_i + \Delta_i \).

Ad [1.4] Regardless of a specific case 0, A, B, or AB and regardless of which \( i \) we consider, the mapping \( \psi \) needs to be continuous on the interior of \( I_i \), or else the discrete part of the measure \( \mathcal{L}_\# \mu \) would have an additional contribution (but we assumed the collection \( (r_i) \), of atoms of \( \mathcal{L}_\# \mu \) to be complete[2]).

We begin with the case that \( i = 1, \ldots, N - 1 \) (where we do not have to take care of whether we are in case 0, A, B, or AB). If \( \alpha_i + \Delta_i + 1 = \alpha_i \), then there actually is nothing to show as \( \psi \) is constant. Consider then \( \alpha_i + \Delta_i + 1 < \alpha_i \). Continuity at the right boundary of \( I_i \), i.e. at \( r = r_{i+1} \) follows from left-continuity of \( \psi \) at this point (Step [1.2]). Monotonicity is inherited from Step [1.1] and it remains to prove surjectivity. But this is clear from the construction of the intervals \( I_i \) and \( J_i \). It holds that \( \lim_{r \searrow r_i} \psi_i(r) = \alpha_i \) (Step [1.3]) and \( \psi(r_{i+1}) = \alpha_{i+1} + \Delta_i \). Hence by continuity, the map \( \psi_i \) is surjective.

In the case that \( i = 0 \), the first subcase (i.e. \( \psi_0 : \{ \inf \mathcal{L} \} \to \{ 1 \} \)) is again trivial and there is nothing we need to prove. We consider the case where \( \inf \mathcal{L} < r_1 \) and \( \alpha_1 + \Delta_1 < 1 \). We see that \( \psi(\inf \mathcal{L}) = \mu(\{ z \in \Omega : \mathcal{L}(z) \geq \inf \mathcal{L} \}) = 1 \), also \( \psi(r_1) = \alpha_1 + \Delta_1 \). Again by continuity, \( \psi_0 \) needs to be surjective.

The case \( i = N \) is shown “anti-parallely” to \( i = 0 \).

Ad [1.5] Follows directly from definition.

Step 2. Set \( \alpha \in [0, 1] \) and define \( r_\alpha := \inf \{ r \in \mathbb{R} : \psi(r) \leq \alpha \} \). Then:

2.1. For any \( r < r_\alpha \) we have \( \psi(r) > \alpha \).

---

[2] In the sense of “not missing an entry”, not in the sense of completeness with respect to convergence.
2.2. If $\alpha \in A_i = [\alpha_i, \alpha_i + \Delta_i)$ for some $i$, then $r_\alpha = r_i$ and $\psi(r_\alpha) = \psi(r_i) = \alpha_i + \Delta_i$.

2.3. If $\alpha \not\in \bigcup_i [\alpha_i, \alpha_i + \Delta_i)$, then $\psi(r_\alpha) = \alpha$.

2.4. $\alpha \mapsto r_\alpha$ is non-increasing: For any $\alpha \geq \beta \geq 0$, we have $r_\alpha \leq r_\beta$.

2.5. If $r_\alpha > r_i$, then $\alpha < \alpha_i$, or equivalently, $\alpha \geq \alpha_i$, then $r_\alpha \leq r_i$.

2.6. $r_\alpha = \sup\{r \in \mathbb{R} : \psi(r) > \alpha\}$.

2.7. If we define $\tilde{r}_\alpha = \inf\{r \in \mathbb{R} : \phi(r) \leq \alpha\} = \sup\{r \in \mathbb{R} : \phi(r) > \alpha\}$ with $\phi$ from 1.5, then $\tilde{r}_\alpha = r_\alpha$. As $\phi(r) = X(r)$, we also have $\tilde{r}_\alpha = \mathcal{L}(\alpha)$ (as defined in section 2).

Proof of Step 2.2 Ad 2.2. If $r < r_\alpha$, then by definition of $r_\alpha$ we have $\psi(r) > \alpha$.

Ad 2.3. First, we know that $\psi(r_i) = \alpha_i + \Delta_i > \alpha$ and for $r < r_i$, by monotonicity, $\psi(r) > \alpha$ as well. Secondly, $\lim_{r \searrow r_i} \psi(r) = \alpha_i \leq \alpha$ and thus, again by monotonicity, $\psi(r) \leq \alpha$ for any $r \geq r_i$. This proves that

$$\{r \in \mathbb{R} : \psi(r) \leq \alpha\} = (r_i, \infty).$$

Taking the infimum, we obtain $r_\alpha = r_i$ and thus $\psi(r_\alpha) = \psi(r_i) = \alpha_i + \Delta_i$.

Ad 2.3. In this case, we are in one of the following scenarios:

- $\alpha \in [\alpha_{i+1} + \Delta_{i+1}, \alpha_i) = J_i$ for $i = 1, \ldots, N - 1$ with $\alpha_{i+1} + \Delta_{i+1} < \alpha_i$.

The case $\alpha_{i+1} + \Delta_{i+1} = \alpha_i$ is impossible because then $\alpha = \alpha_i$ and then we are actually in the territory of 2.2.

By surjectivity (Step 1.4) of $\psi|_{J_i}$ onto $J_i$, the set $\{r \in I_i : \psi(r) = \alpha\}$ is non-empty. Now for any $r' \in \{r \in I_i : \psi(r) < \alpha\}$ and $r'' \in \{r \in I_i : \psi(r) = \alpha\}$, we have $r' \geq r''$. This means that $r_\alpha = \inf\{r \in I_i : \psi(r) \geq \alpha\} = \inf\{r \in I_i : \psi(r) = \alpha\}$. In particular, $r_\alpha \in \overline{I_i}$ with $\overline{I_i}$ being the closure of $I_i$ and there is a minimizing sequence $r^{(n)} \searrow r_\alpha \in I_i$ with $\psi(r^{(n)}) = \alpha$.

Now we claim that $r_\alpha > r_i$. This is true because assuming $r_\alpha = r_i$, then $\psi(r^{(n)}) = \alpha \xrightarrow{n \to \infty} \alpha$ (as the trivial limit of a constant sequence) but by Step 1.3 we know that $\psi(r^{(n)}) \to \alpha_i > \alpha$, which is impossible. Hence, $r_\alpha \in I_i$ and $\psi(r_\alpha) = \alpha$.

- Case 0, A, i.e. $r_N < \sup \mathcal{L}$ and $\alpha \in [0, \alpha_N) = J_N$.

This is proven in the same way as the previous case.

- Case 0, B, i.e. $r_1 > \inf \mathcal{L}$ and $\alpha \in [\alpha_1 + \Delta_1, 1] = J_0$

Even easier because there is no “forbidden” right boundary in the
domain $J_0$.

- Case A, AB, i.e. $r_1 = \inf \mathcal{L}$ and $\alpha \in \{1\} = J_0$, which means that $\alpha = 1$.

Again, this is trivial because in this case obviously $\{r \in I_0 = \{\inf \mathcal{L}\} : \psi(r) \leq \alpha = 1\} = \{\inf \mathcal{L}\}$ and the infimum of this set is its one and only element, hence $r_\alpha = r_1 = \inf \mathcal{L}$ and by direct evaluation, $\psi(r_\alpha) = \alpha_1 + \Delta_1 = 1$.

Ad 2.4: Follows immediately from definition.
Ad 2.5: We prove the equivalent form: Let $\alpha \geq \alpha_i$. We know that $r_\alpha = r_i$ (from Step 2.2) and thus by setting $\beta = \alpha_i$ in Step 2.4, we get $r_\alpha \leq r_\alpha = r_i$. Note that this is not just a corollary from Step 2.1 but a stronger statement in the case that $r = r_i$: Step 2.1 combined with 2.2 just yields the weaker statement $\alpha < \psi(r_i) = \alpha_i + \Delta_i$.

Ad 2.6: This is elementary.
Ad 2.7: This is due to the fact that $\phi$ is the right-continuous version of $\psi$. Take the supremization sequence $r_n$ in the alternative definition of $r_\alpha$ in 2.6, i.e. $\psi(r_n) > \alpha$, $r_n \not\rightarrow r_\alpha$ and $\psi(r) \leq \alpha$ for any $r > r_\alpha$. We can choose a subsequence such that $r_n \neq r_i$ for any $i$. Then $\phi(r_n) = \psi(r_n) > \alpha$ and $r_n \not\rightarrow r_\alpha$. Because $\phi(r)$ is less or equal than $\psi$, we also have $\psi(r) \leq \alpha$ for any $r > r_\alpha$. This means that $r_n$ is also a supremization sequence for $\tilde{r}_\alpha$. \hfill $\square$

**Step 3.** Set $\alpha \in [0,1]$. Then\footnote{The reader might wonder why we defined $\psi$ as we did, i.e. with an $\geq$ instead of an $>$. The central identity $\mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) \leq \alpha$ assumes that $\mathcal{L}(x) < r_\alpha$. Then by Step 2.1, $\alpha < \psi(\mathcal{L}(x)) = \mu(\{z \in \Omega : \mathcal{L}(z) \geq \mathcal{L}(x)\}) = \mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) + \mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\}) \leq \alpha + \mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\})$, hence $\mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\}) > 0$.}

\[ \{x \in \Omega : \mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) \leq \alpha\} = \{x \in \Omega : \mathcal{L}(x) \geq r_\alpha\}. \]

**Proof of Step 3** \footnote{The reader might wonder why we defined $\psi$ as we did, i.e. with an $\geq$ instead of an $>$. The central identity $\mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) \leq \alpha$ assumes that $\mathcal{L}(x) < r_\alpha$. Then by Step 2.1, $\alpha < \psi(\mathcal{L}(x)) = \mu(\{z \in \Omega : \mathcal{L}(z) \geq \mathcal{L}(x)\}) = \mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) + \mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\}) \leq \alpha + \mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\})$, hence $\mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\}) > 0$.}

“$\subseteq$”: Take $x \in \Omega$ such that $\mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) \leq \alpha$. Assume that $\mathcal{L}(x) < r_\alpha$. Then by Step 2.1, $\alpha < \psi(\mathcal{L}(x)) = \mu(\{z \in \Omega : \mathcal{L}(z) \geq \mathcal{L}(x)\}) = \mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) + \mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\}) \leq \alpha + \mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\})$, hence $\mu(\{z \in \Omega : \mathcal{L}(z) = \mathcal{L}(x)\}) > 0$.}
This is only possible if $L(x) = r_i$ with $r_i = L(x) < r_\alpha$ by assumption. The last statement yields (using Step 2.5) that $\alpha < \alpha_i$. But this means that
\[
\alpha \geq \mu(\{z \in \Omega : L(z) > L(x)\}) = \mu(\{z \in \Omega : L(z) > r_i\}) = \alpha_i > \alpha,
\]
a contradiction.

\text{“$\supseteq$”: Let $x \in \Omega$ such that $L(x) \geq r_\alpha$. We need to distinguish two cases.}

Case I: $L(x) = r_\alpha$. Then
\[
\mu(\{z \in \Omega : L(z) > L(x)\}) = \mu(\{z \in \Omega : L(z) > r_\alpha\}) = \mu(\{z \in \Omega : L(z) \geq r_\alpha\}) - \mu(\{z \in \Omega : L(z) = r_\alpha\}) = \psi(r_\alpha) - \mu(\{z \in \Omega : L(z) = r_\alpha\}) = \begin{cases} 
(\alpha_i + \Delta_i) - \Delta_i \leq \alpha, & \text{for } \alpha \in [\alpha_i, \alpha_i + \Delta_i] \text{ for some } i \\
\alpha - " \geq 0" \leq \alpha, & \text{else} 
\end{cases}
\]

Case II: $L(x) > r_\alpha$.
Then by definition of $r_\alpha$, $\alpha \geq \psi(L(x)) = \mu(\{z \in \Omega : L(z) \geq L(x)\}) \geq \mu(\{z \in \Omega : L(z) > L(x)\})$.
This means that in each case, $\mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha$ which concludes the proof of inclusion. 

\text{Step 4. If } \alpha \not\in \bigcup_i [\alpha_i, \alpha_i + \Delta_i], \text{ then } \mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\}) = \alpha.

If $\alpha \in \bigcup_i [\alpha_i, \alpha_i + \Delta_i], \text{ then } \mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\}) = \alpha_i + \Delta_i$.

\text{Proof of Step 4} We calculate (for the first equality we use Step 3)
\begin{align*}
\mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\}) & = \mu(\{x \in \Omega : L(x) > r_\alpha\}) \\
& = \psi(r_\alpha) \\
& = \begin{cases} 
\alpha, & \text{if } \alpha \not\in \bigcup_i [\alpha_i, \alpha_i + \Delta_i] \\
\alpha_i + \Delta_i, & \text{if } \alpha \in \bigcup_i [\alpha_i, \alpha_i + \Delta_i] 
\end{cases}
\end{align*}

This concludes the proof of Lemma 1.
Proposition 1. Let $\Omega \subseteq \mathbb{R}^d$ with $d \in \mathbb{N}$ and $\mathcal{L} : (\Omega, \mathcal{B}^d) \to (\mathbb{R}, \mathcal{B})$ with Borel $\sigma$-algebra $\mathcal{B}^d$ on $\Omega$ and $\mathcal{B}$ on $\mathbb{R}$. Let further $(\Omega, \mathcal{B}^d, \mu)$ be a probability space. We assume that we know the discrete part of $\mathcal{L} \# \mu$ as in Lemma 1, i.e.

\[
\mathcal{L} \# \mu(\{r_i\}) = \mu(\{z \in \Omega : \mathcal{L}(z) = r_i\}) = \Delta_i \\
\mathcal{L} \# \mu((r_i, \infty)) = \mu(\{z \in \Omega : \mathcal{L}(z) > r_i\}) = \alpha_i.
\]

We assume the $\alpha_i$ to be ordered, i.e. $\alpha_1 < \alpha_1 + \Delta_1 \leq \alpha_2 < \cdots$. If there are only $N$-many jumps, we set $\alpha_{N+1} = 1$.

Then, $\Phi : \Omega \to [0,1]$ given by

\[
\Phi(x) = X(\mathcal{L}(x)) = \int_{\mathcal{L}(z) > \mathcal{L}(x)} d\mu(z)
\]

is distributed according to the law

\[
\sum_i \Delta_i \cdot \delta_{\alpha_i} + \text{Unif} \left([0,1] \setminus \bigcup_i [\alpha_i, \alpha_i + \Delta_i]\right),
\]

where $\delta_p$ denotes the Dirac measure at $p$ and $\text{Unif}(I)$ is a uniform distribution on the set $I$.

Proof. The law of $\Phi \in [0,1]$ is the cumulative distribution function of the push-forward $\Phi \# \mu$ given by

\[
\mu \circ \Phi^{-1}([0, \alpha]) = \mu(\{x \in \Omega : \Phi(x) \in [0, \alpha]\}) \\
= \mu \left( \left\{ x \in \Omega : \int_{\mathcal{L}(z) > \mathcal{L}(x)} d\mu(z) \leq \alpha \right\} \right) \\
= \mu \left( \{x \in \Omega : \mu(\{z \in \Omega : \mathcal{L}(z) > \mathcal{L}(x)\}) \leq \alpha\} \right) \\
= \begin{cases} 
\alpha_i + \Delta_i & \text{for } \alpha \in [\alpha_i, \alpha_i + \Delta_i) \\
\alpha & \text{else.}
\end{cases}
\]

This is exactly the cumulative distribution function of the measure

\[
\sum_i \Delta_i \cdot \delta_{\alpha_i} + \text{Unif} \left([0,1] \setminus \bigcup_i [\alpha_i, \alpha_i + \Delta_i]\right).
\]

$\square$
Corollary 1. Let $\Omega \subseteq \mathbb{R}^d$ with $d \in \mathbb{N}$ and $L : (\Omega, \mathcal{B}^d) \to (\mathbb{R}, \mathcal{B})$ with Borel $\sigma$-algebra $\mathcal{B}^d$ on $\Omega$ and $\mathcal{B}$ on $\mathbb{R}$. We assume $L$ is bounded from above and below and that $\mu \circ L^{-1}$ has no discrete atoms, i.e. for any $r \in [\inf L, \sup L]$, we have $\mu(\{x \in \Omega : L(x) = r\}) = 0$.

Then if $x \in \Omega$ is a random variable distributed according to $\mu$, then $\Phi(x)$ is uniformly distributed on $[0, 1]$.

Proof. Follows immediately from proposition[1] \qed

Proposition 2. With the assumptions and definitions of proposition[1] we have that

$$\int_0^1 \tilde{L}(t) \, d\Phi_{#\mu}(t) = \int_0^1 \tilde{L}(t) \, dt. \quad (10)$$

Proof. First we recall from proposition[1] that

$$\Phi_{#\mu} = \sum_i \Delta_i \cdot \delta_{\alpha_i} + \text{Unif} \left( [0, 1] \setminus \bigcup_i [\alpha_i, \alpha_i + \Delta_i] \right).$$

Thus,

$$\int_0^1 \tilde{L}(t) [d\Phi_{#\mu}(t) - dt] = \sum_i \left[ \tilde{L}(\alpha_i) \cdot \Delta_i - \int_{\alpha_i}^{\alpha_i + \Delta_i} \tilde{L}(t) \, dt \right]$$

$$= \sum_i \int_{\alpha_i}^{\alpha_i + \Delta_i} \left[ \tilde{L}(\alpha_i) - \tilde{L}(t) \right] \, dt$$

Figure 9: $\tilde{L}(\alpha_i) \cdot \Delta_i = \int_{\alpha_i}^{\alpha_i + \Delta_i} \tilde{L}(t) \, dt$
By characterization 2.7 in the proof of Lemma 1, we know that for \( t \in [\alpha_i, \alpha_i + \Delta_i] \),
\[
\tilde{L}(t) = \tilde{r}_t = r_t = r_i = \tilde{L}(\alpha_i).
\]
This means that the integrand is always zero and thus
\[
\int_0^1 \tilde{L}(r) [d\Phi \mu(r) - dr] = 0.
\]
For a visualization of this fact see figure 9.
We have shown that although \( \Phi \mu \) is not the uniform measure on \([0, 1]\),
integration of \( \tilde{L} \) with respect to both measures yields the same value. Now we need to validate the second dubious equality, i.e. (3).

4 \( \mu \) -almost-surely \( \hat{L} \circ X \circ \mathcal{L} = \mathcal{L} \)
We defined the generalized inverse for \( X \) in Section 2 as \( \hat{L}(\xi) = \sup \{ \lambda \in \text{im} \mathcal{L} : X(\lambda) > \xi \} \). There, we already indicated that neither \( \hat{L}(X(\lambda)) = \lambda \) for \( \lambda \in \mathbb{R} \) holds nor does \( \hat{L}(X(\mathcal{L}(x))) = \mathcal{L}(x) \) \( \forall \lambda \in \Omega \). This is demonstrated in the following with specific examples. We then continue to show that at least \( \hat{L}(X(\mathcal{L}(x))) = \mathcal{L}(x) \) for \( \mu \)-almost-all \( x \). Although more restrictive, this meaning of “\( \hat{L} \) is the inverse of \( X \)” is all we need in order for the equality in Equation (3)
\[
\int_\Omega \mathcal{L}(x) d\mu(x) = \int_\Omega \hat{L}(X(\mathcal{L}(x))) d\mu(x)
\]
to hold.

4.1 \( \hat{L}(X(\lambda)) \neq \lambda, \lambda \in \mathbb{R} \)
It can be easily demonstrated that the identity \( \hat{L}(X(\lambda)) \neq \lambda \) does not hold for \( \lambda \in \mathbb{R} \) in general. In particular, if \( X \) has a plateau, this fails: Given \( \lambda_1 \neq \lambda_2 \) with \( X(\lambda_1) = X(\lambda_2) \), there is no way to define an inverse \( \hat{L} \) because \( X \) is not injective. See Figure 10 for a concrete example of a likelihood function \( \mathcal{L} \) such that \( X \) is not injective. As is apparent, jumps in the range of \( \mathcal{L} \) are a problem (if they are in the support of \( \mu \)).

The counterexample seems to suggest that only \( \lambda \notin \text{range} \mathcal{L} \) are a problem and that we could at least expect the identity \( \hat{L} \circ X = \text{id} \) to hold on the range of \( \mathcal{L} \), i.e. \( \hat{L}(X(\mathcal{L}(x))) = \mathcal{L}(x) \), but as the next section reveals, this is not true, either.
It is not clear how to define $\tilde{L}(\xi)$ in these ranges. Therefore, its inverse is not even defined especially at the jumps (row 2).
4.2 \( \tilde{L}(X(L(x))) \neq L(x) \) for all \( x \in \Omega \)

We think about when and how \( \tilde{L}(X(L(x))) = L(x) \) is true:

\[
\tilde{L}(X(L(x))) = \sup \{ \lambda \in \text{im} L : X(\lambda) > X(L(x)) \}
\]

As for \( \lambda \geq L(x) \) we have \( X(\lambda) \leq X(L(x)) \), we can ignore any \( \lambda \geq L(x) \) in the specification of the set on the right hand side, hence

\[
= \sup \{ \lambda \in \text{im} L : \lambda < L(x) \text{ and } X(\lambda) > X(L(x)) \}
\]

By definition, \( X(\lambda) = \mu(\{ z \in \Omega : L(z) > \lambda \}) \) and thus \( X(\lambda) - X(L(x)) = \mu(\{ z \in \Omega : L(x) \geq L(z) > \lambda \}) \) (this is due to the elementary fact that for sets \( A \subset B \) and a measure \( \nu \), we have \( \nu(B) - \nu(A) = \nu(B \setminus A) \)).

\[
= \sup \{ \lambda \in \text{im} L : \lambda < L(x) \text{ and } \mu(\{ z \in \Omega : L(z) \in (\lambda, L(x]) \}) > 0 \}
\]

We need to discern two cases now:

**Case I: For all \( \lambda \in \text{im} L \) with \( \lambda < L(x) \), we have that \( \mu(\{ z \in \Omega : L(z) \in (\lambda, L(x]) \}) > 0 \).**

As for \( \lambda = L(x) \) and thus \( (\lambda, L(x]) = \emptyset \), we clearly have \( \mu(\{ z \in \Omega : L(z) \in (\lambda, L(x]) \}) = 0 \), we know that

\[
\sup \{ \lambda \in \text{im} L : \lambda < L(x) \text{ and } \mu(\{ z \in \Omega : L(z) \in (\lambda, L(x]) \}) > 0 \}
\]

\[
= \sup \{ \lambda \in \text{im} L : \lambda < L(x) \}
\]

\[
= L(x),
\]

i.e. indeed

\[
\tilde{L}(X(L(x))) = L(x).
\]

**Case II: (Opposite of Case I)** There exists a \( \lambda^* \in \text{im} L \), i.e. \( \lambda^* = L(x^*) \), \( x^* \in \Omega \), such that \( \lambda < L(x) \) with the property that

\[
\mu(\{ z \in \Omega : L(z) \in (\lambda^*, L(x]] \}) = 0,
\]

then of course

\[
\sup \{ \lambda \in \text{im} L : \lambda < L(x) \text{ and } \mu(\{ z \in \Omega : L(z) \in (\lambda, L(x]) \}) > 0 \} \leq \lambda^*
\]
It is not clear how to define \( \tilde{L}(\xi) \) at this point.

and thus

\[
\tilde{L}(X(L(x))) \leq \lambda^* = L(x^*) < L(x),
\]

i.e. the desirable property \( \tilde{L}(X(L(x))) = L(x) \) does not hold here. It is not difficult to construct a concrete example for this case, see Figure 11.

4.3 \( \tilde{L}(X(L(x))) = L(x) \) for \( \mu \)-almost-all \( x \)

We saw that the \( \tilde{L}(X(L(x))) = L(x) \) does not hold for all \( x \in \Omega \) but that it is violated if for a given \( x \in \Omega \), there exists a \( \lambda^* = L(x^*) < L(x) \) such that

\[
\mu(\{z \in \Omega : L(z) \in (\lambda^*, L(x)]\}) = 0.
\]
We will call such points \( x \in \Omega \) problematic points in this context. We are interested in whether the set of problematic points has vanishing \( \mu \)-measure. If that is the case, then the identity \( \hat{\mathcal{L}}(X(\mathcal{L}(x))) = \mathcal{L}(x) \) still holds \( \mu \)-almost-everywhere. The following lemma shows that this is the case.

**Lemma 2.** The set of problematic points has vanishing \( \mu \)-measure.

**Proof.** The set of problematic points is defined as

\[
P := \left\{ x \in \Omega : \text{There is a } \lambda^* = \mathcal{L}(x^*) < \mathcal{L}(x) \text{ such that } \mu(\{z \in \Omega : \mathcal{L}(z) \in (\lambda^*, \mathcal{L}(x)]\}) = 0 \right\}
\]

Every problematic point \( x_p \in P \) “generates” an interval \([\lambda_p^*, \mathcal{L}(x_p)]\) which has vanishing \( \mathcal{L}_{\#} \mu \)-measure, i.e. \( \mu(\{z \in \Omega : \mathcal{L}(z) \in (\lambda_p^*, \mathcal{L}(x_p)]\}) = 0 \), or \( \mathcal{L}_{\#} \mu((\lambda_p^*, \mathcal{L}(x_p)]) = 0 \). The (possibly uncountable) union of those intervals \( I_p \) over all \( x_p \) becomes a countable union of intervals,

\[
\bigcup_p I_p = \bigcup_{n \in \mathbb{N}} (a_n, b_n] \cup \bigcup_{m \in \mathbb{N}} (c_m, d_m),
\]

where \( a_n \) and \( c_m \) are limits of certain subsequences \( \lambda_{p_m}^{(n)} \) and \( b_n \) and \( d_m \) are limits of certain subsequences \( \mathcal{L}(x_p^{(n)}) \), respectively. This follows from Lemma 5. By continuity from below for \( \mathcal{L}_{\#} \mu \), we also have that \( \mathcal{L}_{\#} \mu((a_n, b_n]) = \mu(\{z \in \Omega : \mathcal{L}(z) \in (a_n, b_n]\}) = 0 \), for any \( n \in \mathbb{N} \), and the same for the intervals \( (c_m, d_m) \). As countable sums of terms with value 0 are 0, we obtain \( \mathcal{L}_{\#} \mu(\bigcup_p I_p) = 0 \).

Now note that for every \( x_p \in P \), by construction, \( \mathcal{L}(x_p) \in I_p \), which in turn means that

\[
P \subseteq \left\{ x \in \Omega : \mathcal{L}(x) \in \bigcup_p I_p \right\} = \mathcal{L}^{-1}\left( \bigcup_p I_p \right).
\]

Finally, we obtain \( \mu(P) \leq \mu \circ \mathcal{L}^{-1}\left( \bigcup_p I_p \right) = \mathcal{L}_{\#} \mu \left( \bigcup_p I_p \right) = 0 \).

### 5 Counterexamples to nested sampling

In this section, we use realistic examples to demonstrate that nested sampling does not necessarily compute the correct evidence even though the conditions

---

4Here, \( p \) is an index which is not necessarily countable but which indexes the whole family of points \( x_p \in P \).
of theorem [1] hold. In order to keep things simple, we investigate a linear 1D example
\[ y = x + \epsilon, \quad (11) \]
with one datum \( y \in \mathbb{R} \), the unknown parameter \( x \in \mathbb{R} \) and some measurement noise \( \epsilon \in \mathbb{R} \). The following representative scenarios for the measure and the measurement noise resulting in a specific likelihood function are

- **Scenario 1**: Uniform measure and Gaussian likelihood – 1D
- **Scenario 2**: Uniform measure and truncated Gaussian likelihood (case A) – 1D
- **Scenario 3**: Gaussian measure and clipped off Gaussian likelihood (case B) – 2D

Note that the parameter space \( \Omega \) has been restricted in the first two scenarios to ensure that the prior has full support.

The plot of the function \( \tilde{\mathcal{L}}(\xi) \) for each scenario is shown in Figure 12. As we already know from [17], cliffs are not problematic. In contrast, plateaus mean that \( \Phi_{\#\mu} \) is not uniform. Thus, the correctness of nested sampling cannot be guaranteed because the original estimation of \( X_k = e^{-\frac{1}{n}} \) for \( X(\mathcal{L}(x)) \) at iteration \( k \) is not valid any longer. We will see in our experiments that a plateau indeed leads to a completely wrong calculation of the evidence via nested sampling. However, in case of very narrow plateaus in \( \tilde{\mathcal{L}} \) the original nested sampling can still work sufficiently well.

The nested sampling algorithm has been implemented in Matlab 2019a according to the algorithm given in [16]. For generating new active samples from a randomly chosen one, the slice sampling algorithm from Matlab (slicesample) was modified such that no stepping out is performed and the likelihood constraint is satisfied. The algorithm nested sampling is probabilistic and yield slightly varying estimations of the logevidence for repeating
runs. Thus, we generate 6000 independent runs with \( n = 500 \) active samples for each scenario. Each run is iterated until the stopping criterion \( c \cdot n \cdot H < k \), with \( c = 1000 \), \( H \) being the information or negative entropy (for details see [16]) and \( k \) the number of iteration in nested sampling, is fulfilled or the relative change in the logevidence drops below \( c_p = 10^{-7} \). Nested sampling provides additionally an uncertainty of the estimated logevidence which is given by \( \pm \sqrt{\frac{H}{n}} \) [17]. In all examples, the initial set of active samples is randomly drawn according to the measure.

### 5.1 Uniform measure and Gaussian likelihood – 1D

In case of additive Gaussian noise \( (\xi \sim \mathcal{N}(\mu_\xi, \sigma^2) \) with mean \( \mu_\xi \in \mathbb{R} \) and standard deviation \( \sigma > 0 \), the likelihood function is also Gaussian and, therefore, this scenario is a common example for an experimental setting. We restrict the parameter space to \( \Omega = [-3, 3] \) and set the datum \( y = 0.5 \).

In detail, the likelihood function \( L : \Omega \to \mathbb{R} \) for \( \mu_\xi = 0 \) and \( \sigma = 1 \) is given by

\[
L(x) = \frac{1}{\sqrt{2\pi}} e^{- \frac{(0.5-x)^2}{2}}
\]

We further assume a uniform measure \( \mu \) on \( \Omega \) whose density \( \pi : \Omega \to \mathbb{R}_+ \) is given by

\[
\pi(x) = \frac{1}{6}
\]

Note that the prior has full support on \( \Omega \). The evidence is the integral over the unnormalized posterior \( \tilde{P}(x \mid y) : \Omega \to \mathbb{R}_+ \)

\[
\tilde{P}(x \mid y) = L(x) \cdot \pi(x) = \frac{1}{6} \cdot \frac{1}{\sqrt{2\pi}} e^{- \frac{(0.5-x)^2}{2}}
\]

because it is the normalization constant of the unnormalized posterior and, thus,

\[
Z_1 = \int L(x) \cdot \pi(x) \, dx = \frac{1}{6} \int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{- \frac{(0.5-x)^2}{2}} \, dx
\]

\[
= \frac{1}{6} \left( \Phi_{\text{norm}}(y + 3) - \Phi_{\text{norm}}(y - 3) \right)
\]

\[
\approx 0.1656
\]

\[
\log Z_1 \approx -1.7982
\]
with $\Phi_{\text{norm}}$ being the cumulative density function of the standard normal distribution. A visualization of the likelihood and the density of the uniform measure are given in Figure 13. There is also a boxplot of the estimated log-evidence values given. The median ($-1.7941$) and the mean of the estimated log-evidences ($-1.7952$) are close to the true value. The true log-evidence lies within one uncertainty in 50.48% of all runs and within two uncertainties in 81.90%. The spread of the estimated log-evidence close around the true value is also obvious in the according histogram.

In conclusion, nested sampling can be said to work perfectly here because the function $\tilde{\mathcal{L}}$ has neither a cliff nor a plateau.

5.2 Uniform measure and truncated Gaussian likelihood – 1D

Let us again assume additive noise of a Gaussian form but this time we know that there exists an upper bound for the measurement noise. This situation can be motivated by scenarios where there is a known maximal measurement error. In the following, we set this bound on the error to 1 and the parameter space $\Omega = [-3, 3]$. Again, the datum is set to $y = 0.5$. Then, the additive measurement noise is distributed according to a truncated normal distribution with mean 0 and variance 1 resulting in the likelihood function $\mathcal{L} : \Omega \to \mathbb{R}_+$

\[
\mathcal{L}(x) = \begin{cases} 
C \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.5-x)^2}{2}}, & x \in (-0.5, 1.5) \\
0, & \text{else}
\end{cases}
\]

with $C = \frac{1}{\Phi_{\text{norm}}(1) - \Phi_{\text{norm}}(-1)}$ (see also Figure 14).

As the algorithm nested sampling is usually implemented using the log-likelihood function, we set $\mathcal{L}(x) = 10^{-323}$ for $x \notin (-0.5, 1.5)$ in the implementation to enable the calculation of $\log \mathcal{L}(x)$. We choose as previously a uniform measure $\mu$ on $\Omega$. The unnormalized posterior is accordingly given by

\[
\tilde{P}(x \mid y) = \mathcal{L}(x) \cdot \pi(x) = \begin{cases} 
\frac{1}{6} \cdot C \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(0.5-x)^2}{2}}, & x \in (-0.5, 1.5) \\
0, & \text{else}
\end{cases}
\]

because in our case $x \in (-0.5, 1.5)$ is more restrictive than $x \in \Omega = [-3, 3]$. 

29
Problem Setting

measure $\mu(x)$

function $L(x)$

$-3 \leq y = 0.5 \leq 3$

$\alpha$

$\mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\})$

Cumulative distribution function of $\mu \circ \Phi^{-1}$

$\mu(\{x \in \Omega : \mu(\{z \in \Omega : L(z) > L(x)\}) \leq \alpha\})$

Run logevidence

Histogram of log $Z_1$

analytical logevidence

Figure 13: Scenario 1: uniform measure and Gaussian likelihood. (row 1) Problem setting and cumulative distribution function of $\mu \circ \Phi^{-1}$. The uniform assumption of the pushforward measure is fulfilled. (row 2) Boxplot and Histogram of 6000 estimated logevidence values together with the analytical logevidence. The boxplot is additionally supplemented with 13 representative runs with their estimates of the logevidence with uncertainty bars, plotted in comparison to the analytical logevidence.
Therefore, the evidence $Z_2$ is

$$Z_2 = \int_{-3}^{3} \tilde{P}(x \mid y) \, dx$$

$$= \frac{1}{6} C \cdot \int_{-0.5}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.5-x)^2}{2}} \, dx$$

$$= \frac{1}{6} \cdot \frac{1}{\Phi_{\text{norm}}(1) - \Phi_{\text{norm}}(-1)} \{\Phi_{\text{norm}}(1) - \Phi_{\text{norm}}(-1)\}$$

$$= \frac{1}{6}$$

$$\approx 0.1667$$

$$\log Z_2 \approx -1.7916$$

nested sampling overestimates the logevidence resulting in a mean of $-1.3582$ for 6000 conducted runs (Figure [14]) as also visualized by the evolution of the logevidence of one representative nested sampling run (Figure [15]). Additionally, the evolution of the logevidence over iterating is awkward at the beginning. This results from the fact that the likelihood function plateaus at its minimum value. However, neither a higher number of active samples nor a variation of the value for $L$ in case of $x \notin (-0.5, 1.5)$ improves the estimation of the logevidence (data not shown) which indicates that the cause is more complex. In fact, it seems that the violation of the uniform assumption for the pushforward measure $\mu \circ \Phi^{-1}$ results in the overestimation of the logevidence as the estimation of $X_k = e^{-\frac{k}{n}}$ for $X(L(x))$ at iteration $k$ is not valid any longer. The correct logevidence is neither within one uncertainty nor within two in all runs.

5.3 Gaussian measure and clipped off Gaussian likelihood – 2D

Now we assume that the instrument adds measurement noise

- from a range equally likely, e.g. $\|\xi\| \leq \xi^*$
- less and less likely the further away from this interval.

An example for this scenario, is a hypothetical measurement instrument doing quantization in a specific range, e.g. rounding of measurements. This leads to complete ignorance about the true value within the range of quantization. We set $\Omega = \mathbb{R}^2$ for the parameter space and assume the datum
Problem Setting

\[ \text{measure } \mu(x) \]

\[ \text{function } L(x) \]

\[ y = -3 \]

\[ y - 1 \]

\[ y = 0 \]

\[ y + 1 \]

\[ 3 \]

Cumulative distribution function of \( \mu \circ \Phi^{-1} \)

Visualization of uncertainty of \( \log Z_2 \)

Boxplot and Histogram of 6000 estimated log evidence values together with the analytical log evidence. The boxplot is additionally supplemented with 13 representative runs with their estimates of the log evidence with uncertainty bars, plotted in comparison to the analytical log evidence.

Figure 14: Scenario 2: uniform measure and truncated Gaussian likelihood. (row 1) Problem setting and cumulative distribution function of \( \mu \circ \Phi^{-1} \). The uniform assumption of the pushforward measure is not fulfilled. (row 2) Boxplot and Histogram of 6000 estimated log evidence values together with the analytical log evidence. The boxplot is additionally supplemented with 13 representative runs with their estimates of the log evidence with uncertainty bars, plotted in comparison to the analytical log evidence.
Figure 15: Evolution of the log-evidence during iteration of nested sampling in scenario 2. For scaling reasons we defined $L(x) = 10^{-10}$ for $x \notin (-0.5, 1.5)$ just for this plot.

Figure 16: Scenario 3: clipped off Gaussian likelihood in 2 dimensions and sketch of cumulative distribution function of $\mu \circ \Phi^{-1}$. The uniform assumption of the pushforward measure is not fulfilled.

$y = (0.25, -0.25)^\top$. The likelihood $L : \mathbb{R}^2 \to \mathbb{R}_+$ is then of the following form

$$L(x) = \min \left\{ \frac{1}{2 \pi \sigma_1 \sigma_2} e^{-\frac{(0.25-x_1)^2}{2 \sigma_1^2}} - \frac{(-0.25-x_2)^2}{2 \sigma_2^2} \right\}$$

with $C = \frac{1}{2 \pi \sigma_1 \sigma_2} e^{-\frac{\sigma_1^2}{2 \sigma_1^2} - \frac{\sigma_2^2}{2 \sigma_2^2}}$ and variances $\sigma_1^2 = 0.3$ and $\sigma_2^2 = 0.4$ which is a clipped off Gaussian likelihood (Figure 16). We assume a prior Gaussian measure $\mu(x)$ with mean $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$ and density $\pi(x)$. The evidence $Z_3$ of the posterior is approximated as the mean of the likelihood of $10^7$ samples from $\mu$. Its value is approximately $Z_3 \approx 0.0279$ and $\log Z_3 \approx -3.5790$, respectively. Indeed, the algorithm nested sampling fails
to calculate the evidence in this setting independently of the number of active samples. This results in an underestimated mean logevidence value \( \log Z_3 = -4.5834 \) and \( \log Z_3 = -4.5861 \) for 500 and 5000 active samples, respectively, and 6000 independent runs. This can be also clearly seen from the boxplot and histograms of the logevidence values in Figure 17. As soon as all active samples have the same likelihood which is \( C \), no better active sample in the sense of a higher likelihood value exists. Thus, the algorithm must terminate even though the logevidence has not converged yet (see Figure 18).

6 Robust implementation of nested sampling

In the previous section, we saw that the nested sampling algorithm has difficulties with proper calculation of the evidence if the likelihood has plateaus. Using the insights deduced from the proof of theorem 1, we develop an adapted nested sampling method such that it can deal with plateaus in the likelihood successfully.

To achieve this, the parameter space \( \Omega \) is decomposed into disjoint subsets: on \( \Omega_c \subseteq \Omega \) (c for “constant”) the likelihood function has plateaus and on \( \Omega_r = \Omega \setminus \Omega_c \) (r for “regular”) it does not. The computation of evidence, mean and variance of the posterior are conducted separately for either subset.

On \( \Omega_c \), integration should be easier because the most difficult integrand (the likelihood) is piecewise constant, although in practice we do not know the shape and measure of \( \Omega_c \) which usually precludes analytical computation. For that reason we present two versions of nested sampling on \( \Omega_c \): Analytical or with numerical integration.

On \( \Omega_r \), we essentially just do original nested sampling, with the only necessary modification being the restriction of the initial accessible prior mass of \( \Omega_r \) which might be less than 1. Afterwards, the results on \( \Omega_r \) and \( \Omega_c \) are merged suitably.

This approach also enables the generation of weighted samples from the posterior. We then apply this more robust implementation of nested sampling to the previous counterexamples and show that evidence, mean and variance are approximated quite satisfactorily.

6.1 Decomposition of parameter space \( \Omega \)

Let us assume that the likelihood has a finite number of plateau heights \( h = 1, \ldots, H, H \in \mathbb{N} \) denoted by \( \mathcal{L}_{h,c} \). We define \( \Omega_{h,c} := \{ x \in \Omega : \mathcal{L}(x) = \mathcal{L}_{h,c} \} \) and set \( \Omega_c = \bigcup_{h=1}^{H} \Omega_{h,c} \).
Uncertainty of $\log Z_3$
(500 active samples)

$\log Z_3$
run

Histogram of $\log Z_3$
(500 active samples)

$\log Z_3$
Frequency

Figure 17: Scenario 3: Gaussian measure and clipped off Gaussian likelihood. (row 1) Boxplot of 6000 estimated logevidence values and representative variation of the logevidence with uncertainty bars compared to the analytical logevidence for 500 and 5000 active samples, respectively. (row 3) Histogram of 6000 estimated logevidence values together with the analytical logevidence for 500 and 5000 active samples, respectively.
We stick with the notation of Lemma 1 by defining \( \Delta_h = \mu(\Omega_{h,c}) \). Note that \( r_i = \mathcal{L}_{i,c} \) but we opted for the slightly more expressive notation \( \mathcal{L}_{i,c} \) for \( i = 1, \ldots, H \) here.

We assume that the plateau heights \( \mathcal{L}_{h,c} \) are known in advance (i.e. we need to be able to decide whether some point \( x \) is an element of \( \Omega_c \) by checking its likelihood value \( \mathcal{L}(x) \)) but we do not necessarily need the numerical value of the \( \Delta_h \).

The likelihood has no plateaus of positive measure on the remaining set \( \Omega_r = \Omega \setminus \Omega_c \). Therefore, nested sampling can be applied on \( \Omega_r \) successfully. As the maximum accessible prior mass of \( \Omega_r \) is \( 1 - \sum_h \Delta_h \) (instead of 1), we use its reduced value also within nested sampling. Nevertheless, the accessible prior volume keeps shrinking by the factor \( e^{\frac{1}{n}} \) from iteration to iteration. The decomposition of \( \Omega \) and the meaning of all introduced quantities is visualised in Figure 19.

### 6.2 Estimation of \( \Delta_h \)

If \( \Delta_h \) cannot be calculated analytically which is likely to be the case in practice, it can be approximated during the generation of the initial set of active samples which are in \( \Omega_r \). If we draw \( N \) samples \( x_i, i = 1, \ldots, N \) from \( \mu \), then \( \Delta_h \) is approximately the ratio of samples within \( \Omega_{h,c} \) (relative frequency), i.e. \( \Delta_h = \frac{n_h}{N} \) with \( n_h \) being the number of samples in \( \Omega_{h,c} \). To derive this and an additional confidence interval to the confidence level \( \gamma \) for \( N \) sufficiently large formally, we introduce the random variable \( X^h \) on \( (\Omega, \mathcal{B}_\Omega) \) with \( X^h_i = \delta_{x_i \in \Omega_{h,c}} \). Then, \( \Delta_h \) is the mean of \( X^h \) with its estimator \( \hat{\Delta}_h = \frac{n_h}{N} \). The variance of \( X^h \) is estimated as \( \hat{s}^2 = \frac{1}{N-1} \sum_{i=1}^N (X^h_i - \hat{\Delta}_h)^2 \).

The confidence interval for \( \Delta_h \) can be derived via Student’s \( t \)-distribution.
Figure 19: Decomposition of $\Omega$ for the robust implementation of nested sampling.

The evidence calculation is decomposed into the two parts on $\Omega_r$ and $\Omega_c$ such that one obtains

$$Z = \int_{\Omega} L(x) \, d\mu(x)$$

$$= \int_{\Omega_r} L(x) \, d\mu(x) + \int_{\Omega_c} L(x) \, d\mu(x)$$

$$\approx Z_r + \sum_h L_{h,c} \cdot \Delta_h$$

(12)
with $Z_r$ denoting the evidence obtained from nested sampling on $\Omega_r$ already considering the maximum accessible prior volume of $1 - \sum_h \Delta_h$ as $\xi_0$. The visualization in terms of $\tilde{L}$ is given in Figure 20. Indeed, the reduction of $\xi_0$ weights automatically the contributions from nested sampling and $\Omega_c$ to $Z$ because the width of the bars in Figure 20(left) becomes rescaled by this factor.

There it becomes obvious that one either has to weight the evidence obtained from original nested sampling with $1 - \sum_h \Delta_h$ in the sense that the width of the bars is shrunk or (what we do) one shrinks the maximum accessible prior mass to $1 - \sum_h \Delta_h$ beforehand. Both approaches are analytically equivalent. However, for the calculation of the mean and variance of the posterior and for the generation of posterior samples the latter approach is more convenient.

In addition to the evidence calculation, nested sampling allows to estimate the mean and the variance of the posterior simultaneously. The corresponding integrals are decomposed analogously to the evidence calculation:

$$m_{\text{tot}} = \frac{1}{Z} \int_{\Omega} x \cdot L(x) \, d\mu(x)$$
$$= \frac{1}{Z} \left[ Z_r \int_{\Omega_r} x \cdot \frac{L(x)}{Z_r} \, d\mu(x) + \int_{\Omega_c} x \cdot L(x) \, d\mu(x) \right]$$
$$\approx \frac{1}{Z} \left[ Z_r \cdot m_r + \sum_h L_{h,c} \cdot \Delta_h \cdot m_{h,c} \right] \quad (13)$$

If we write $Z_{h,c} = L_{h,c} \cdot \Delta_h = \int_{\Omega_{h,c}} L(x) \, d\mu(x)$, then this can be interpreted
as the average of the means on all the set $\Omega_r, \Omega_{h,c}$, $h = 1, \ldots, H$, weighted by their respective contribution to the total evidence. 

$$
\sigma_{\text{tot}}^2 = \frac{1}{Z} \int_{\Omega} x^2 \cdot L(x) \, d\mu(x) - m_{\text{tot}}^2 
$$

$$
= \frac{1}{Z} \left[ Z_r \int_{\Omega_r} x^2 \cdot \frac{L(x)}{Z_r} \, d\mu(x) + \int_{\Omega_c} x^2 \cdot L(x) \right] - m_{\text{tot}}^2 
$$

$$
\approx \frac{1}{Z} \left[ Z_r (\sigma_r^2 + m_r^2) + \sum_h L_{h,c} \cdot \Delta h \cdot \sigma_{h,c}^2 \right] - m_{\text{tot}}^2 \tag{14}
$$

with $m_r$ and $\sigma_r^2$ being the mean and variance obtained from nested sampling on $\Omega_r$ as the following holds:

$$
m_r = \int_{\Omega_r} x \cdot \frac{L(x)}{Z_r} \, d\mu \approx \sum_k w_k x_k \tag{15}
$$

$$
\sigma_r^2 = \int_{\Omega_r} x^2 \cdot \frac{L(x)}{Z_r} \, d\mu - m_r^2 \approx \sum_k w_k (x_k - m_r)^2 \tag{16}
$$

Here, $w_k$ is the weight of sample $x_k$ obtained from the nested sampling step.

Note that $\mu$ is not a measure on $\Omega_r$, but on $\Omega$. Nevertheless, this is already factored in due to the reduced maximum accessible prior mass. The approximations are given e.g. in [16, section 9.4.1]. The derivation of the integrals over $\Omega_{h,c}$ is sketched in Section 6.2.

However, mean and variance are only of limited significance for non-Gaussian posterior. Therefore, we describe how to generate weighted samples from the posterior within the framework. Mean and variance can also be estimated from these weighted samples.

### 6.4 Weighted samples from the posterior

The weights of samples from the posterior obtained via nested sampling are their relative contribution to the evidence [16, section 9.4.1]. Since we have already taken into account the accessible prior volume of $\Omega_r$ in our nested sampling step, the samples $x_k$ in $\Omega_r$ and their weights $w_k$ remain the same. The weight of all samples within $\Omega_{h,c}$ is equal and depends on the height of the likelihood plateau $L_{h,c}$ and on the number of samples within $\Omega_{h,c}$ denoted by $n_h$ as follows

$$
w_{h,c} = \frac{1}{Z} \cdot \frac{\Delta h}{n_h} \cdot L_{h,c} \tag{17}
$$
The weights are already normalized, i.e. \( \sum_k w_k + \sum_h w_{h,c} = 1 \). From these weighted samples, the mean and variance of the posterior can be approximated. This is especially useful if \( \Delta_h, m_{h,c} \text{ and } \sigma_{h,c}^2 \) in Equation \((13) - (14)\) cannot be calculated analytically.

### 6.5 Roadmap of robust implementation

The following summary explains shortly the roadmap for evidence calculation, the generation of weighted samples from the posterior and the approximation of mean and variance of the posterior using the robust implementation.

1. generate \( N \) initial samples \( x_i, i = 1, ..., N \) from measure \( \mu \) and sort them (Section 6.1):
   - \( x_i \in \Omega_r \): sample becomes part of active set for nested sampling
   - \( x_i \in \Omega_{h,c} \): sample used for estimation of \( \Delta_h, m_{h,c}, \sigma_{h,c}^2 \) and generation of weighted samples from posterior

2. calculate evidence
   (a) if \( \Delta_h \) unknown, estimate it via relative frequencies (Section 6.2), i.e. \( \Delta_h = \frac{n_h}{N} \) with \( n_h \) number of samples \( x_i \in \Omega_{h,c} \)
   (b) perform nested sampling on \( \Omega_r \) considering the maximum accessible prior mass, i.e. \( \xi_0 = 1 - \sum_h \Delta_h \) (instead of \( \xi_0 = 1 \) in original nested sampling) and obtain \( Z_r \) as depicted in Figure 20
   (c) compute \( Z_c = \sum_h L_{h,c} \cdot \Delta_h \)
   (d) calculate evidence \( Z \approx Z_r + Z_c \) according to Equation \((12)\)

3. generate weighted samples from posterior separated by domains:
   (a) \( \Omega_r \): use dead samples from nested sampling with their weights
   (b) \( \Omega_c \): use initial samples in \( \Omega_{h,c} \) and assign their weights according to Equation \((17)\)

4. estimate mean and variance of posterior via one of the two methods
   - derivation from weighted samples from posterior
   - direct calculation (Section 6.3)
     (a) derive \( m_r \) and \( \sigma_r^2 \) from weighted dead samples obtained from nested sampling (Equation \((15) - (16)\)
(b) if mean \( m_{h,c} \) and variance \( \sigma^2_{h,c} \) of \( \Omega_{h,c} \) unknown, estimate them from their representative samples

(c) calculate mean and variance of posterior according to Equations (13) – (14)

We remark that the robust implementation becomes the original nested sampling as long as the likelihood has no plateaus of positive measure on \( \Omega \), namely \( \sum_h \Delta_h = 0 \).

6.6 Application of robust implementation

In this section, we apply the robust implementation of nested sampling to scenario 2 and 3 from Section 5 where the original nested sampling implementation failed. The stopping criterion was augmented to \( c = 2000 \).

6.6.1 Uniform measure and truncated Gaussian likelihood – 1D

The parameter space \( \Omega = [-3, 3] \) is decomposed into \( \Omega_{1,c} = [-3, -0.5] \cup [1.5, 3] \) and \( \Omega_c = (-0.5, 1.5) \). In this example we have only one plateau with \( \Delta_1 = \frac{2}{3} \) and height \( L_{1,c} = 10^{-323} \). We draw 50 and 500 samples, respectively, from \( \mu \) and conduct 6000 runs to average the evidence, mean and variance. The results are given in Table 1. Evidence, mean and variance approximate their correct value satisfactorily, see also Figure 21. Note that there can be seen that the number of active samples for nested sampling varies for each run in accordance with Section 6.2. The mean and variance from the calculation in Section 6.3 and from the weighted samples is equivalent in terms of histograms and mean values. The histogram of weighted samples from one representative run fits the normalized posterior Figure 22.

6.7 Gaussian measure and clipped off Gaussian likelihood – 2D

In application, the robust implementation does not require the explicit definition of \( \Omega_{h,c} \). In fact, it is sufficient to define it via the likelihood values \( L_{h,c} \). It holds \( x \in \Omega_{h,c} \) if and only if \( L(x) = L_{h,c} \) and numerically \( \| L(x) - L_{h,c} \| \leq \epsilon \) with sufficiently small \( \epsilon > 0 \). In this scenario, \( \Delta_{1,c}, m_{1,c} \) and \( \sigma_{h,c} \) are estimated via samples from the Gaussian measure \( \mu \). However, the latter ones are only required for the calculation of mean and variance of the posterior as described in Section 6.3 and not for their estimation from the weighted samples as in Section 6.4. Table 2 summarizes the estimated evidence, mean and variance averaged from 6000 independent runs with 50 and 500 samples.
Figure 21: Scenario 2 with robust implementation: Uniform measure and truncated Gaussian likelihood. Histograms of 6000 runs for 500 samples drawn from $\mu$. (row 1) The evidence ranges around the numerical evidence. The number of active samples in nested sampling varies for each run. (row 2) Mean and variance calculated directly according to Section 6.3. (row 3) Mean and variance estimated via weighted samples from posterior.
Table 1: Scenario 2: Comparison of analytical results and estimations obtained from the robust implementation (means and standard deviation of 6000 runs). Mean and variance are estimated additionally from weighted samples.

|                  | evidence | mean   | variance   |
|------------------|----------|--------|------------|
| **numerical solution** | 0.1667   | 0.5    | 0.2911     |
| **50 samples**   | 0.1772 ± 0.0336 | 0.5017 ± 0.1328 | 0.2667 ± 0.0669 |
|                  | direct calculation | weighted samples |
| **500 samples**  | 0.1678 ± 0.0107 | 0.5005 ± 0.0419 | 0.2891 ± 0.0215 |
|                  | direct calculation | weighted samples |

Figure 22: Scenario 2 with robust implementation: Histogram of weighted samples from posterior obtained from robust implementation of nested sampling with 5000 samples from the prior together with normalized posterior function (red curve).
Figure 23: Scenario 3 with robust implementation: Gaussian measure and clipped off Gaussian likelihood. Histogram of evidence values from 6000 runs for (left) 50 samples drawn from $\Omega$ and (right) 500 samples.

drawn from $\Omega$ at the very beginning of the robust implementation. Figure 23 shows that computation of the evidence can be done already with very high accuracy for 500 samples drawn from $\Omega$. An increase of number of samples will decrease the uncertainty of the evidence estimation, i.e. the width of the histogram would shrink further. This can be already seen from the difference between the histograms of the evidences for 50 and 500 samples from $\Omega$. This results from the decreased uncertainty within the application of nested sampling on $\Omega_r$ and the reduced uncertainty of estimation of $\Delta_{h_i}$. In the original nested sampling algorithm, the number of active samples steers the uncertainty of approximations whereas in the robust implementation the number of samples drawn from $\Omega$ at the very beginning is crucial.

A Appendix

We recall the following fact from measure theory.

**Lemma 3.** Any open subset $U \subseteq \mathbb{R}$ can be written as an at most countable union of disjoint open intervals, i.e.

$$U = \bigcup_{n \in \mathbb{N}} (c_n, d_n)$$

with $d_n > c_n$ and $(c_n, d_n) \cap (c_m, d_m) = \emptyset$ for $m \neq n$.  

44
Table 2: Scenario 3: Comparison of results from evidence estimation and Metropolis-Hastings and estimations from the robust implementation. Numerical mean and variance have been approximated via the random walk Metropolis-Hastings algorithm. Evidence, mean and variance were approximated via the robust implementation (means and standard deviation of 6000 runs with 50 and 500 samples drawn independently from Ω). Mean and variance are estimated additionally from weighted samples.

|                  | evidence | mean      | variance   |
|------------------|----------|-----------|------------|
| **numerical solution** | 0.0279   | 0.4928    | 0.5609     |
|                  |          | −0.3482   | 0.7234     |
| **50 samples**   |          |           |            |
| direct calculation | 0.5007 ± 0.1161 | 0.5684 ± 0.0677 |
|                  |          | −0.3519 ± 0.1458 | 0.7333 ± 0.0773 |
| weighted samples  | 0.5008 ± 0.1161 | 0.5552 ± 0.0643 |
|                  |          | −0.3520 ± 0.1458 | 0.7180 ± 0.0736 |
| **500 samples**  |          |           |            |
| direct calculation | 0.4944 ± 0.0369 | 0.5614 ± 0.0199 |
|                  |          | −0.3486 ± 0.0457 | 0.7248 ± 0.0225 |
| weighted samples  | 0.4944 ± 0.0369 | 0.5601 ± 0.0198 |
|                  |          | −0.3486 ± 0.0457 | 0.7233 ± 0.0224 |
Lemma 4. Let $\Pi$ be a set (possibly uncountable) of indices and $\{I_p\}_{p \in \Pi}$ be a (possibly uncountable) family of intervals of the form $I_p = (a_p, b_p)$ with $b_p > a_p$ for any index $p \in \Pi$. Then

$$\bigcup_{p \in \Pi} I_p = \bigcup_{n \in \mathbb{N}} (c_n, d_n)$$

(18)

with $d_n > c_n$ and $(c_n, d_n) \cap (c_m, d_m) = \emptyset$ for $m \neq n$, i.e. the uncountable union simplifies to a countable union of open intervals.

Furthermore, for any $n \in \mathbb{N}$, there is a sequence of indices $\{p_m^{(1)}\}_m \subset \Pi$ such that $c_n = \lim_{m \to \infty} a_{p_m^{(1)}}$ and a sequence of indices $\{p_m^{(2)}\}_m \subset \Pi$ such that $d_n = \lim_{m \to \infty} b_{p_m^{(2)}}$.

Proof. An arbitrary union (even uncountable) of open sets is itself open, thus $U = \bigcup_{p \in \Pi} I_p$ is open. Then we can apply Lemma 3 to $U$ which immediately yields the expression (18). Now we define $\Pi_i$ as the subset of indices such that

$$\bigcup_{p \in \Pi_i} I_p = (c_i, d_i).$$

This is possible because each interval $I_p$ can only be a subset of exactly one of the disjoint sets $(c_n, d_n)$. Now clearly $c_i = \inf_{p \in \Pi_i} a_p$ and $d_i = \sup_{p \in \Pi_i} b_p$. Then $\{p_m^{(1)}\}_m$ can be chosen as the infimization sequence for $c_i$ and analogously for $d_i$.

Lemma 5. Let $\{I_p\}_{p \in \Pi}$ be a (possibly uncountable) family of intervals of the form $I_p = (a_p, b_p]$ with $b_p > a_p$ for any $p \in \Pi$. Then

$$\bigcup_{p \in \Pi} I_p = \bigcup_{n \in \mathbb{N}} J_n$$

where each $J_n$ is either $J_n = (c_n, d_n]$ or $(c_n, d_n]$ with $c_n, d_n$ being the same numbers as in Lemma 4, i.e. the uncountable union simplifies to a countable union of open and half-open intervals.

Proof. From the proof of Lemma 4, we know that

$$\bigcup_{p \in \Pi_i} I_p = (c_n, d_n) \cup \bigcup_{p \in \Pi_i} \{b_p\}.$$

Now there are two cases to be distinguished: If there is a $p^* \in \Pi_i$ such that $b_{p^*} = \sup_{p \in \Pi_i} b_p = d_n$, then $\bigcup_{p \in \Pi_i} I_p = (c_n, b_{p^*}] = (c_n, d_n]$. If that is not the case, i.e. if for any $p \in \Pi_i$, we have $b_p < d_n$, then $\bigcup_{p \in \Pi_i} I_p = (c_n, d_n)$. $\square$
Acknowledgement

Both authors want to thank Maria Neuss-Radu for fruitful discussion. D.S. likes to acknowledge the support by the Bavarian Equal Opportunities Sponsorship — Realisierung von Chancengleichheit von Frauen in Forschung und Lehre (FFL) — Realization Equal Opportunities for Women in Research and Teaching.

References

[1] Stuart Aitken and Ozgur E. Akman. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. BMC Systems Biology, 7(1):72, 07 2013.

[2] Nicolas Chopin and Christian Robert. Contemplating evidence: properties, extensions of, and alternatives to nested sampling. Technical report, Citeseer, 2007.

[3] Arnaud de La Fortelle. A study on generalized inverses and increasing functions part i: generalized inverses. 2015.

[4] Richard Dybowski, Trevelyan J. McKinley, Pietro Mastroeni, and Olivier Restif. Nested sampling for bayesian model comparison in the context of salmonella disease dynamics. PLOS ONE, 8(12):1–17, 12 2013.

[5] Paul Embrechts and Marius Hofert. A note on generalized inverses. Mathematical Methods of Operations Research, 77(3):423–432, 2013.

[6] F. Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. Mon. Not. R. Astron. Soc., 384:449–463, 2008.

[7] 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.

[8] F Feroz, MP Hobson, E Cameron, and AN Pettitt. Importance nested sampling and the multinest algorithm. arXiv preprint arXiv:1306.2144, 2013.

[9] Hans-Otto Georgii. Stochastics. De Gruyter, Berlin, Boston, 2013.
[10] Edward Higson, Will Handley, Michael Hobson, and Anthony Lasenby. Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. Statistics and Computing, 29(5):891–913, 2019.

[11] Edward Higson, Will Handley, Mike Hobson, Anthony Lasenby, et al. Sampling errors in nested sampling parameter estimation. Bayesian Analysis, 13(3):873–896, 2018.

[12] Thomas Kämpke and Franz Josef Radermacher. The generalized inverse of distribution functions. In Income Modeling and Balancing, pages 9–28. Springer, 2015.

[13] Pia Mukherjee, David Parkinson, and Andrew R. Liddle. A nested sampling algorithm for cosmological model selection. The Astrophysical Journal, 638(2):L51–L54, 01 2006.

[14] David Parkinson and Andrew R. Liddle. Bayesian model averaging in astrophysics: a review. Statistical Analysis and Data Mining, 6(1):3–14, 01 2013.

[15] Nick Pullen and Richard J. Morris. Bayesian model comparison and parameter inference in systems biology using nested sampling. PLOS ONE, 9(2):1–11, 02 2014.

[16] D. S. Sivia and J. Skilling. Data Analysis: A Bayesian Tutorial. Oxford University Press, 2nd edition, 2006.

[17] John Skilling et al. Nested sampling for general bayesian computation. Bayesian analysis, 1(4):833–859, 2006.