Neutron star parameter estimation from a NICER perspective

Riley, T.E.

Publication date
2019

Document Version
Other version

License
Other

Link to publication

Citation for published version (APA):
Riley, T. E. (2019). *Neutron star parameter estimation from a NICER perspective*.

B

X-PSI: A prototype open-source package for neutron star X-ray Pulsation Simulation and Inference

B.1 X-ray frequency evolution

We are primarily concerned with sources whose signals are stable over many pulsations. Suppose first that the X-ray frequency is non-evolving and the source of radiation co-rotates (i.e., is coherent) with the stellar surface. It follows that phases computed based on a (radio derived) timing solution $\tilde{\tau} \mapsto \phi$ may be assumed to be valid X-ray phases—at least up to a global phase shift—such that $\phi_X \equiv \phi$ and thus $\varphi_X \equiv \varphi = \phi - \lfloor \phi \rfloor$. This is the case for rotation-powered X-ray pulsars.

However, it is possible to expand scope in a simplistic manner to consider sources whose X-ray pulsation frequency—due to the global behavior of radiating surface material—evolves with respect to the coordinate time of the chart globally foliating the spacetime. The rotational phase function $\phi(t)$ is then not equivalent to the X-ray phase function given by

$$\phi_X(t) = \phi_{X,0} + \frac{1}{2\pi} \int_0^t \omega_X(t') dt' \tag{B.1}$$

where $\omega_X(t)$ is the coordinate angular frequency of X-ray pulsations, and the zero phase $\phi_{X,0} \coloneqq \phi_X(t = 0)$. Thus, if a map $\tilde{\tau} \mapsto \phi$ is applied to arrival events based on a radio timing solution (Section 3.2.6), the derived event phases are not valid X-ray phases. However, if indeed the X-ray frequency $\omega_X(t)$ evolves in time *relative to the stellar rotation frequency* $\omega(t)$, it is possible to statistically model such evolution given *non-folded* event arrival data with phases ϕ defined by the rigid rotation of the stellar source matter (i.e., the rotation of the surface). The X-ray frequency evolution can be treated in the generative model of the event arrival data via a parameterized invertible map $\phi_X \mapsto \phi$. One then replaces the phase variable ϕ in Section 3.2.7 through Section 3.2.10 with the variable ϕ_X, and similarly φ goes to φ_X, and so on. Resultantly, the map $\phi \mapsto \varphi_X$ is parameterized.

An assumption that *may* be verifiable via a preliminary Fourier analysis is the following:[1] the $\omega_X(t)$ identified as the fundamental mode of pulsation via a discrete fast Fourier transform deviates only slightly from (a harmonic of) $\omega(t)$. As an example, the frequency of flux oscillations during a thermonuclear burst drifts by $\mathcal{O}(1)$ Hz over many cycles from the nominal spin frequency. The map $\phi \mapsto \varphi_X$ can thus be parametrized and an *unbinned* likelihood function (Section 3.2.9) can be defined which uses an economical[2] pulsation signal.

In order to explain the simple forms of time-evolution supported in *X-PSI*, we decouple the evolution of the surface radiation field into the following components: (i) the instantaneous radiation field as a function of spatial coordinates in a *chart* that co-rotates with some mode of asymmetry in the surface photosphere and is instantaneously a Schwarzschild chart; (ii) the coordinate angular frequency of that mode of asymmetry with respect to the Schwarzschild chart; and (iii) the coordinate rotation frequency of the fluid light-element surface material, assumed to be equivalent to that of the solid surface. Regarding component (ii): evolution is permitted, but the phase-delay along rays is evaluated given some coordinate angular frequency that is invariant over one revolution.

Regarding component (i): we consider *general* time-evolution of the radiation field in the co-rotating chart as beyond the scope of *X-PSI* for the foreseeable future; nevertheless, for certain restricted forms of time-evolution there is support. For example, one can implement a custom extension module that modifies the definition of the economically computed pulsation to be an average over some arbitrary number of cycles; crucially, the discretization of the surface needs to be both sufficiently high-resolution *and* time-independent (beyond rigid rotation), such that intensity of radiation from a given surface element can be simply summed over revolutions and subsequently propagated to a distant observer (refer to the appendix of Bogdanov et al. 2019, submitted to ApJLb, for related comments on discretization).

As a toy example, let us suppose that we have to contend with an $\mathcal{O}(1)$ Hz frequency drift of the type detected in the *tails* of thermonuclear burst oscillation signals (Galloway et al. 2008; Watts 2012; Bilous & Watts 2018), but an accurate (toy) assumption is that the local comoving surface radiation field is non-evolving between Schwarzschild time hyperslices—i.e., non-evolving besides evolution at fixed spatial point on the surface due to rotation. Let the frequency drift be explained by a global fluid mode that propagates azimuthally, at evolving angular frequency, in a prograde or retrograde sense relative to the bulk periodic rotation of the radiating fluid. The coordinate angular frequency of the X-ray oscillations would affect the photon flux function but would not control the Lorentz transformations at the surface. Therefore, in this limit, one could argue that the effect is small and neglect frequency evolution for simulation of one rotational pulse—i.e., component (ii) above. The pulse could then be invoked for describing some segment of the burst over many rotational cycles: phase-fold according to $\phi \mapsto \varphi_X$ and evaluate an unbinned likelihood function. Whether such a simple model could perform adequately would need to be tested.

[1] Although this can be notoriously difficult in practice for some sources to which such analysis is particularly applicable (Bilous & Watts 2018).

[2] Refer to Section 3.2.7.

B.2 Marginalization of the default background parameters

In Section 3.2.10.3 we defined a default background model for phase-folded, binned likelihood functions. The implementation in *X-PSI* is enabled by the background parameters being *fast*, and by the tractability of likelihood function marginalization. Marginalization reduces the dimensionality of the sampling space to a non-prohibitive number. In this appendix we detail how marginalization is performed accurately but rapidly for each likelihood function evaluation.

B.2.1 Overview

The joint probability *mass* of the data conditional on a vector of model parameters $(\boldsymbol{\theta}, \boldsymbol{B})$ is given by

$$p(\{\boldsymbol{d}_i\}_{i=1,\ldots,I} \mid \boldsymbol{\theta}, \{B_i\}) = \prod_{i,k} p(d_{ik} \mid \boldsymbol{\theta}, B_i), \tag{B.2}$$

where $i = 1, \ldots, I$ denotes detector channels, and $k = 1, \ldots, K$ enumerates a sequence of phase intervals, each of which has an associated number d_{ik} of counts; there is assumed to be *zero* statistical covariance between pairs of d_{ik}, $\forall(i,k)$. Note that contrary to the main body of this paper, in this appendix we identify $\boldsymbol{\theta}$ as any continuous parameters that are *not* marginalized over but form a dimension of the sampling space; these parameters necessarily include the target parameters θ_S, and potentially instrument response parameters as discussed in Section B.3.2. Given Equation (3.34) and assuming the phase intervals $\boldsymbol{\varphi}_k$ are all of equal length $1/K$, let

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{B}) &\coloneqq -2\ln L(\boldsymbol{\theta}, \boldsymbol{B}) \\ &= -2\sum_{i,k} d_{ik} \ln\left[s_{ik}(\boldsymbol{\theta}) + B_i\left(\sum_{\ell}\Delta t_{\ell}\right)\int_{\boldsymbol{\varphi}_k} d\varphi\right] - s_{ik}(\boldsymbol{\theta}) - B_i\left(\sum_{\ell}\Delta t_{\ell}\right)\int_{\boldsymbol{\varphi}_k} d\varphi \\ &= -2\sum_{i,k} d_{ik} \ln\left[\frac{T_{\text{exp}}}{K}\left(\bar{s}_{ik}(\boldsymbol{\theta}) + B_i\right)\right] - \frac{T_{\text{exp}}}{K}\left(\bar{s}_{ik}(\boldsymbol{\theta}) + B_i\right),\end{aligned} \tag{B.3}$$

up to a known constant, where $T_{\text{exp}} \coloneqq \sum_{\ell}\Delta t_{\ell}$ is the total exposure time over all observing intervals, and $\bar{s}_{ik}(\boldsymbol{\theta})$ is the phase-averaged, expected count rate in $(i,k)^{th}$ interval, contributed by the target source.

B.2.2 Dangers of the default background model

First, as discussed in Chapter 4, the relative contribution from surface hot regions may be unconstrained *a priori*. The phase-invariant background terms are thus in principle permitted to absorb some phase-invariant contribution from the *real* surface hot regions—supposing that the physical description of the target source is a close approximation to reality. If the model does not perform adequately *a posteriori*, it should be clear that model development is necessary; if, on the other hand, the model does perform adequately, concerns about relative contribution are arguably not too meaningful because the ground truth is unknown.

Second, depending on the target, there may be a coherent source component with a dependence on the principal phase φ that is not captured by the target integrals (see Section 3.2.3) nor these background parameters. Discussion on this issue may be found in Miller & Lamb (2015) for bursting sources, and can also be applicable in the context of isolated X-ray pulsars. Such a component would need to be generated by radiating material in the local vicinity of the star that corotates with it. Examples of potentially important pulsed signals from accretion-powered pulsars that would not captured by this background model include: (i) accretion disc reflection; and (ii) off-surface emission from an accretion funnel that is coupled to the rotating magnetic field of the star. Notably, these components are also dependent on exterior spacetime parameters.

For non-accreting sources such as rotation-powered X-ray pulsars, off-surface emission may be unimportant. It may however be reasonable to consider a pulsed signal generated by uniform surface emission exterior to closed hotter regions. The signal would be anti-phased relative to the combined signal from the closed hotter regions, and is the simplest additional component to calculate in *X-PSI* (at slightly increased computational cost).[3] The importance of such a component is dependent on instrument sensitivity to softer X-rays than typically emitted by the hotter regions primarily targeted.

B.2.3 Marginalization

We are interested in marginalization of the target (joint posterior) distribution $\pi(\boldsymbol{\theta},\boldsymbol{B}) := p(\boldsymbol{\theta},\boldsymbol{B}\,|\,\boldsymbol{d})$:[4]

$$\pi(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta},\boldsymbol{B})d\boldsymbol{B} \propto p(\boldsymbol{\theta})\int L(\boldsymbol{\theta},\boldsymbol{B})p(\boldsymbol{B})d\boldsymbol{B}; \tag{B.4}$$

we denote the joint prior distribution as $p(\boldsymbol{\theta},\boldsymbol{B})$. In order to draw samples directly from the posterior $\pi(\boldsymbol{\theta})$ we therefore require the likelihood as a function of $\boldsymbol{\theta}$, marginalized over the $\boldsymbol{B}$-subspace:

$$L(\boldsymbol{\theta}) := \int L(\boldsymbol{\theta},\boldsymbol{B})p(\boldsymbol{B})d\boldsymbol{B}; \tag{B.5}$$

hereafter we refer to this function as the *background-marginalized* likelihood function or simply as the *marginal* likelihood function. The joint prior $p(\boldsymbol{B})$ is assumed to be separable

[3] It can be efficiently computed via modification of the local radiative specific intensity within the boundaries of the hotter regions. The signal from the hot regions can then be linearly combined with the phase-invariant signal from the surface when ignoring the hotter regions.

[4] We are working in the context of a single model $\mathcal{M}_{\text{default}}$ with a continuous associated joint parameter space $(\boldsymbol{\theta},\boldsymbol{B})$, so for brevity we omit the conditional arguments for the model (including any prior information).

and diffuse—the latter being relative to the conditional likelihood function of $\boldsymbol{B}$ given $\boldsymbol{\theta}$.

Let us consider the likelihood function given by Equation (B.3) and identify the expected number of counts in the $(i, k)^{th}$ interval as

$$c_{ik}(\boldsymbol{\theta}, B_i) = \frac{T_{\exp}}{K} \left[\bar{s}_{ik}(\boldsymbol{\theta}) + B_i\right]. \tag{B.6}$$

Due to the nature of the generative model for the photon count data—where the background parameter B_i is unique to the sampling distribution of the data subset in the i^{th} channel—the marginalization operation is separable over the I channels:

$$L(\boldsymbol{\theta}) \propto \int p(\boldsymbol{B}) \left[\prod_{i,k} c_{ik}^{d_{ik}}(\boldsymbol{\theta}, B_i) e^{-c_{ik}}\right] d\boldsymbol{B} = \prod_i \int p(B_i) \left[\prod_k c_{ik}^{d_{ik}}(\boldsymbol{\theta}, B_i) e^{-c_{ik}}\right] dB_i. \tag{B.7}$$

B.2.3.1 Analytical marginalization

We first review analytical and semi-analytical approximations to $\boldsymbol{B}$-subspace marginalization. Let the global maximum in the log-likelihood function be

$$\mathcal{L}_0 = -2 \ln L_0 = -2 \max_{(\boldsymbol{\theta},\boldsymbol{B})} \left[\ln L(\boldsymbol{\theta}, \boldsymbol{B})\right], \tag{B.8}$$

such that the parameter vector that maximizes the likelihood function is given by $\boldsymbol{y}_0 := (\boldsymbol{\theta}_0, \boldsymbol{B}_0)$. A local second-order expansion of the log-likelihood function in the $\boldsymbol{B}$-subspace (a $\boldsymbol{\theta}$-hyperslice through the joint space of all model parameters) about some point $\boldsymbol{y} := (\boldsymbol{\theta}, \boldsymbol{B}')$[5] is given by (Taylor & Kitching 2010)

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{B}) \approx \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{B}') + \delta\boldsymbol{B}^T \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{B}}\right|_{\boldsymbol{y}} + \frac{1}{2}\delta\boldsymbol{B}^T \left.\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{B}^2}\right|_{\boldsymbol{y}} \delta\boldsymbol{B} + \mathcal{O}(\|\delta\boldsymbol{B}\|^3), \tag{B.9}$$

such that under the assumption of local Gaussianity of $L(\boldsymbol{\theta}, \boldsymbol{B})$, its expectation with respect to a joint flat (improper) prior density $p(\boldsymbol{B})$, gives the log-marginal-likelihood function (see the appendix of Taylor & Kitching 2010)

$$\mathcal{L}(\boldsymbol{\theta}) \approx \underbrace{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{B}') - \frac{1}{2} \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{B}}\right|_{\boldsymbol{y}}^T \left.\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{B}^2}\right|_{\boldsymbol{y}}^{-1} \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{B}}\right|_{\boldsymbol{y}}}_{\approx -2 \ln \max_{\boldsymbol{B}} [L(\boldsymbol{\theta},\boldsymbol{B})]} + \mathrm{Tr} \ln \left(\left.\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{B}^2}\right|_{\boldsymbol{y}} \right), \tag{B.10}$$

[5] Where in general $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ as required, and $\boldsymbol{B}'$ need not necessarily be an accurate approximation to the background vector $\boldsymbol{B}_0$ that maximizes the global likelihood function in the full model parameter space Taylor & Kitching (2010), provided the conditional likelihood function on the (nuisance) $\boldsymbol{B}$-subspace does not exhibit appreciable departures from Gaussianity in the peak region.

where the approximation becomes exact if the *conditional* likelihood function on the $\boldsymbol{B}$-subspace is exactly Gaussian.[6] If the parameters $\boldsymbol{\theta}$ were negligibly correlated with the parameters $\boldsymbol{B}$ (over some subdomain of the full parameter space where the joint prior density is finite), and vector $\boldsymbol{B}'$ maximizes the likelihood function for some arbitrary vector $\boldsymbol{\theta}'$ in the peak region, the second term in Equation (B.10) involving first-order derivatives with respect to $\boldsymbol{B}$ is in practice close to vanishing $\forall\boldsymbol{\theta}$: $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{B}') \approx -2\ln\max_{\boldsymbol{B}}\left[L(\boldsymbol{\theta},\boldsymbol{B})\right]$. Note that this second term is restricted to being either positive or zero. The third term will in general always be negative to account for the correction to the (approximate) $\ln\max_{\boldsymbol{B}}\left[L(\boldsymbol{\theta},\boldsymbol{B})\right]$ by probabilistic averaging of the assumed Gaussian likelihood function on the $\boldsymbol{B}$-subspace with respect to the nuisance joint prior distribution on that subspace.

Applying this approximation to the likelihood function given by Equation (B.3):

$$\mathcal{L}(\boldsymbol{\theta}) \approx -2\sum_{i,k} d_{ik}\ln\left[\frac{T_{\rm exp}}{K}\left(\overline{s}_{ik}(\boldsymbol{\theta})+B_i\right)\right] - \frac{T_{\rm exp}}{K}\left(\overline{s}_{ik}(\boldsymbol{\theta})+B_i\right) - \frac{1}{2}\frac{\partial\mathcal{L}}{\partial\boldsymbol{B}}\bigg|_{y}^{T}\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{B}^2}\bigg|_{y}^{-1}\frac{\partial\mathcal{L}}{\partial\boldsymbol{B}}\bigg|_{y} + \mathrm{Tr}\ln\left(\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{B}^2}\bigg|_{y}\right). \tag{B.11}$$

All derivatives can be calculated analytically. Let us first calculate the first-order derivatives:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{B}} &= -2\frac{\partial}{\partial\boldsymbol{B}}\left\{\sum_{i,k} d_{ik}\ln\left[\frac{T_{\rm exp}}{K}\left(\overline{s}_{ik}(\boldsymbol{\theta})+B_i\right)\right] - \frac{T_{\rm exp}}{K}\left(\overline{s}_{ik}(\boldsymbol{\theta})+B_i\right)\right\}\\ \therefore\;\frac{\partial\mathcal{L}}{\partial B_i} &= -2\sum_{k} d_{ik}\frac{\partial}{\partial B_i}\ln\left[\frac{T_{\rm exp}}{K}\left(\overline{s}_{ik}(\boldsymbol{\theta})+B_i\right)\right] - \frac{T_{\rm exp}}{K}\\ &= 2T_{\rm exp} - 2\sum_{k} d_{ik}\left(\overline{s}_{ik}(\boldsymbol{\theta})+B_i\right)^{-1}.\end{aligned} \tag{B.12}$$

The conditional likelihood function, as required, has second-order derivatives $\frac{\partial^2\mathcal{L}}{\partial B_j\partial B_i}$ that vanish[7] on $\boldsymbol{\theta}$-hyperslices, where here $j\neq i$ where $j := 1,\dots,I$, whilst

$$\frac{\partial^2\mathcal{L}}{\partial B_i^2} = 2\sum_{k} d_{ik}\left(\overline{s}_{ik}(\boldsymbol{\theta})+B_i\right)^{-2}. \tag{B.13}$$

The second-order derivatives are dependent on $\boldsymbol{B}$, and *all* higher-order derivatives (with respect to non-mixed background parameters) are also finite, so the logarithm of the likelihood function is not quadratic, as expected. Crucially however, there no inflection points on the

[6] Note that for a bounded uniform joint prior on the $\boldsymbol{B}$-subspace, the log-marginal-likelihood function is approximated by Equation (B.10) up to a constant that includes the prior volume in the $\boldsymbol{B}$-subspace (unnecessary for parameter estimation but relevant to evidences).

[7] Note that when *marginalized* over the $\boldsymbol{\theta}$-subspace, the parameters of $\boldsymbol{B}$ may in principle exhibit finite degeneracies due to the effective coupling to the parameters $\boldsymbol{\theta}$, provided there exist finite degeneracies between $\boldsymbol{\theta}$ and $\boldsymbol{B}$.

domain $B_i \in \mathbb{R}_{\geq 0}$. For $n \in \mathbb{N}_{\geq 2}$:

$$\frac{\partial^n \mathcal{L}}{\partial B_i^n} = 2 \times (-1)^n \times (n-1)! \sum_k d_{ik} \left(\bar{s}_{ik}(\boldsymbol{\theta}) + B_i\right)^{-n} ; \tag{B.14}$$

notably, the signs oscillate with order and thus higher-order corrections partially negate in the Taylor expansion—B.9. Further, as the number of counts in the i^{th} channel increases, the expected count numbers within the peak region increase and are raised to the power n in the denominator, such that the higher-order derivatives are rapidly damped with n for large numbers of counts. Asymmetry manifests because there is an additional mode of dependence of the sampling distribution in the $(i, k)^{th}$ interval on B_i, which appears in both the mean and the variance. In the limit that the data consists of large d_{ik} (e.g., for long exposure times), a maximum-likelihood sampling distribution $p(d_{ik} \mid \boldsymbol{\theta}_0, \boldsymbol{B}_0)$ in the $(i, k)^{th}$ interval becomes symmetric, and as does the conditional likelihood function.

Let us calculate the inverse matrix for the likelihood function given by Equation (B.3). The covariance matrix (up to a constant factor) is given by:

$$\left[\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{B}^2}\right]^{-1} = \frac{1}{2}\mathrm{diag}\left[\frac{1}{\sum_k d_{1j}\left(\bar{s}_{1k}(\boldsymbol{\theta}) + B_1\right)^{-2}}, \ldots, \frac{1}{\sum_k d_{Ik}\left(\bar{s}_{Ik}(\boldsymbol{\theta}) + B_I\right)^{-2}}\right], \tag{B.15}$$

and thus letting $\boldsymbol{B}' = (B'_1, B'_2, \ldots, B'_I)$ where $\boldsymbol{y} \triangleq (\boldsymbol{\theta}, \boldsymbol{B}')$:

$$\begin{aligned}
\ln L(\boldsymbol{\theta}) \approx\ & \underbrace{\sum_{i,k} d_{ik} \ln\left[\frac{T_{\mathrm{exp}}}{K}\left(\bar{s}_{ik}(\boldsymbol{\theta}) + B'_i\right)\right] - \frac{T_{\mathrm{exp}}}{K}\left(\bar{s}_{ik}(\boldsymbol{\theta}) + B'_i\right)}_{1^{\mathrm{st}}\ \mathrm{term}} \\
& \underbrace{+\frac{1}{2}\sum_i \left[T_{\mathrm{exp}} - \sum_k d_{ik}\left(\bar{s}_{ik}(\boldsymbol{\theta}) + B'_i\right)^{-1}\right]^2 \left[\sum_k d_{ik}\left(\bar{s}_{ik}(\boldsymbol{\theta}) + B'_i\right)^{-2}\right]^{-1}}_{\mathrm{positive}\ 2^{\mathrm{nd}}\ \mathrm{term}} \\
& \underbrace{-\frac{1}{2}\sum_i \ln \sum_k d_{ik}\left(1 + \frac{\bar{s}_{ik}(\boldsymbol{\theta})}{B'_i}\right)^{-2}}_{\mathrm{negative}\ 3^{\mathrm{rd}}\ \mathrm{term}} + \cancel{\sum_i \ln B'_i - \frac{I \ln 2}{2}},
\end{aligned} \tag{B.16}$$

where the last terms are independent of $\boldsymbol{\theta}$ and are thus unimportant for parameter estimation (conditional on some model) in this approximation. If, $\forall (i, k)$, $B'_i \approx K d_{ik}/T_{\mathrm{exp}}$ and $B'_i \gg \bar{s}_{ik}(\boldsymbol{\theta})$, then the dependence of the third term on $\boldsymbol{\theta}$ is weak, and the data are not informative for constraining $\boldsymbol{B}$ only. However, if $\bar{s}_{ik}(\boldsymbol{\theta}) \sim B'_i$, the dependence of the third term on $\boldsymbol{\theta}$ is stronger if the dependence of the rates $\{\bar{s}_{ik}\}$ on $\boldsymbol{\theta}$ is strong.

Instead of fixing $\forall \boldsymbol{\theta}$ the vector $\boldsymbol{B}'$, an approximation for each B'_i could be invoked. Considering the phase-integrated Poisson process in each channel, approximate conditional likelihood maximization in the i^{th} channel is achieved via

$$\begin{aligned} \sum_{k} d_{ik} &\sim \text{Poisson}\left(T_{\text{exp}}B_i' + \frac{T_{\text{exp}}}{K}\sum_{k}\bar{s}_{ik}(\boldsymbol{\theta})\right) \\ \therefore\ B_i' &\approx \frac{1}{T_{\text{exp}}}\sum_{k} d_{ik} - \frac{1}{K}\sum_{k}\bar{s}_{ik}(\boldsymbol{\theta}) = \frac{K}{T_{\text{exp}}}\left\langle d_{ik}\right\rangle_k - \left\langle \bar{s}_{ik}(\boldsymbol{\theta})\right\rangle_k . \end{aligned} \tag{B.17}$$

Note that in this approximation $B_i' = B_i'(\boldsymbol{\theta})$, and thus the penultimate term of Equation (B.16) is not omitted. Crucially, this approximation can yield *negative* background count rates if the count rate from the star is sufficiently high; in such instances, the conditional likelihood function on the $\boldsymbol{B}$-subspace may not be well-approximated by a Gaussian, in part due to truncation. Otherwise, the approximation promises to improve accuracy of the marginal likelihood function with increasing separation from the global peak on the $\boldsymbol{\theta}$-subspace.

In Miller & Lamb (2015) the likelihood is numerically maximized for each vector $\boldsymbol{\theta}$, which is a naturally more accurate alternative to the maximization term (the second term) in Equation (B.16) if the conditional likelihood function on the $\boldsymbol{B}$-subspace (fixed $\boldsymbol{\theta}$) is *not* well-approximated as Gaussian or vector $\boldsymbol{B}'$ leads to inaccuracies. The numerical maximization should nevertheless be inexpensive relative to target signal computation because the joint sampling distribution of the data is separable over channels, each of which is uniquely associated with a *fast* background rate parameter. One may thus embed some sophisticated low-level one-dimensional maximization routine *within* the marginal likelihood function evaluator, with the routine being executed I times per evaluation given a vector $\boldsymbol{\theta}$ (as can be inferred from the description in Miller & Lamb 2015).

In Miller & Lamb (2015) a different type of approximation is applied: during stochastic mapping of the $\boldsymbol{\theta}$-subspace, additional tracking of the factor difference between the approximate $\boldsymbol{B}$-marginalized likelihood and the $\boldsymbol{B}$-maximized likelihood suggested weak dependence on $\boldsymbol{\theta}$ under the assumption of $\boldsymbol{B}$-subspace *conditional likelihood function Gaussianity*. In terms of the approximation given by Equation (B.10), the third term would need to exhibit an unimportant dependence on $\boldsymbol{\theta}$. In that case, even if sampling noise on the $\boldsymbol{\theta}$-subspace was not a dominant source of error, parameter inferences (in the continuous space defined by some model) would be insensitive to which likelihood is used in practice. However, it is unnecessary for us to rely on this approximation being sufficiently accurate because fast numerical marginalization is tractable.

B.2.3.2 Fast numerical background marginalization

We opt to perform numerical marginalization to circumvent the problem of analytically justifying the invocation of Gaussianity of the conditional likelihood function on the $\boldsymbol{B}$-subspace for general use cases. The resulting algorithm can thus be implemented in the low count-number regime and beyond, and truncation of the marginalization operation at $\boldsymbol{B} = \boldsymbol{0}$ is accurately accounted for.

Numerical integration is, in the same vein as numerical maximization of the conditional likelihood function discussed above, inexpensive relative to target signal computation. It follows that I one-dimensional explicit integrals need to be performed to compute the marginal likelihood of the vector $\boldsymbol{\theta}$, where for each iteration in each channel, the time to reevaluate the conditional likelihood function is small because the background parameters are all *fast*. If, algorithmically, likelihood function evaluation is structured such that varying a B_i for fixed $\boldsymbol{\theta}$ is many orders of magnitude faster than when varying $\boldsymbol{\theta}$, the relative speed difference gives a handle on the maximum number of conditional likelihood function evaluations (over all channels) permitted before numerical marginalization becomes as expensive as target signal calculation. Even if it were the case that marginal likelihood function evaluation was $\mathcal{O}(1)$ times *longer* than likelihood function evaluation, it is arguably justified in order to ensure the number of sampling space dimensions is tractable. A sophisticated low-level one-dimensional *integration* routine can thus be embedded within the marginal likelihood function evaluator, with the routine being executed I times per evaluation given a vector $\boldsymbol{\theta}$.

Let us examine how to perform background marginalization via numerical quadrature in practice. It is necessary to define the prior density distribution $p(B_i)$ on each background subspace. For the purpose of *parameter estimation* given some model, one can in principle define for each B_i an improper prior density distribution—such as uniform on the semi-unbounded domain $\mathbb{R}_{\geq 0}$. Suppose one defines some arbitrarily large, finite upper bound $\mathscr{B}_i$, such that the prior density on the interval $B_i \in [0, \mathscr{B}_i]$ is constant. In the limit that $\mathscr{B}_i \to \infty$, the posterior density distribution is defined due to the integrability of the conditional likelihood function. However, even if the prior density functions for each B_i are proper, if the joint prior density $p(\boldsymbol{B})$ is separable and weak, the model may clearly diverge from being *generative* in the sense that it can generates data that violates some knowledge about the data-generating process (such as instrumental limits), but the model may be adequately *predictive a posteriori* (Gelman et al. 2017).

To enable evidence computation, however, one needs to define finite bounds, because in the limit $\mathscr{B}_i \to \infty$ the evidence is zero for an integrable conditional likelihood function in the i^{th} channel. Evidences will be critically sensitive to this bound because it exerts control over the prior predictive complexity. As we caution in Section 3.2.10.3, however, the prior predictive complexity of a given model is generally always going to be greatly amplified by invoking this default background treatment, even if the bounds $\mathscr{B}_i$ are tailored appropriately to an instrument and the observational fields containing the target source. Therefore, we advise that if one opts to couple this default background treatment with some set of target (and instrument) models, one takes care to apply the same joint prior density distribution on the $\boldsymbol{B}$-subspace in order to eliminate differences in prior predictive complexities due to the background prior choice.

To accurately evaluate the integrand without numerical overflow and with minimal underflow we first consider that

$$\begin{aligned}\ln L(\boldsymbol{\theta}) &= \sum_i \ln \int_{a_i}^{b_i} \exp\left[\sum_k d_{ik}\ln c_{ik} - c_{ik} - \ln d_{ik}!\right] dB_i + \cancel{\sum_i \ln p(B_i)}.\\ &= \sum_i \ln \int_{a_i}^{b_i} \exp\left[\text{const.} + \sum_k d_{ik}\ln c_{ik} - c_{ik}\right] dB_i,\end{aligned} \tag{B.18}$$

where const. $= -\sum_k \ln d_{ik}!$ may simply be precomputed and stored in memory for all likelihood evaluations, and $\ln d_{ik}!$ can be rapidly evaluated for $d_{ik} \lesssim \mathcal{O}(100)$ using hard-coded tables, and via series truncation for larger values—we use the GSL library. We can further manipulate the integrand by writing

$$\begin{aligned}\ln L(\boldsymbol{\theta}) &= \sum_i \ln \int_{a_i}^{b_i} \exp\left[\text{const.} + \sum_k d_{ik}\ln c_{ik} - c_{ik}\right] dB_i\\ &= \sum_i \text{const.} + \max_{B_i}\left[\sum_k d_{ik}\ln c_{ik} - c_{ik}\right]\\ &\qquad + \ln \int_{a_i}^{b_i} \exp\left[\sum_k (d_{ik}\ln c_{ik} - c_{ik}) - \max_{B_i}\sum_k (d_{ik}\ln c_{ik} - c_{ik})\right] dB_i\\ &= \text{const.} + \sum_i \max_{B_i}\left[\sum_k d_{ik}\ln c_{ik} - c_{ik}\right]\\ &\qquad + \ln \int_{a_i}^{b_i} \exp\left[-T_{\text{exp}}B_i + \sum_k (d_{ik}\ln c_{ik}) - \max_{B_i}\left(-T_{\text{exp}}B_i + \sum_k d_{ik}\ln c_{ik}\right)\right] dB_i,\end{aligned} \tag{B.19}$$

and const. $= -\sum_{ik} \ln d_{ik}!$ is thus not necessary unless the fully marginal likelihood is desired. The utility of writing the integrand in this form[8] is that underflow is circumvented; overflow is not a problem in general even working with Equation (B.18), however underflow occurs away from peak regions in each background dimension and the numerical marginal likelihood function becomes noisy, possibly inducing artificial local modes at very small likelihoods. We desire a smooth logarithmic function of the likelihood. Underflow occurs due to computer exponentiation of large negative numbers. In Equation (B.19) however, we ensure the exponent is unity at some point within the integral domain, meaning that the important contributions to the marginalization operation are accurately evaluated. We proceed to discuss the maximiza-

[8] This is merely the integral form of the standard *logsumexp* manipulation.

tion operation, which is performed via numerical iteration, given a fiducial background vector and analytical first and second order partial derivatives of the logarithm of the likelihood function.

In order to robustly capture the peak in the conditional likelihood function, one needs to ensure the applied adaptive quadrature routine initially divides the integration domain into some sufficiently large number of intervals. The integral limits must be specified appropriately: one can base these limits on the data $\{d_{ik}\}$ and the numerical rates $\{\overline{s}_{ik}(\boldsymbol{\theta})\}$. The maximum in conditional likelihood function satisfies

$$\frac{\partial \mathcal{L}}{\partial B_i} = 2T_{\text{exp}} - 2\sum_k d_{ik}\left(\overline{s}_{ik}(\boldsymbol{\theta}) + B_i\right)^{-1} = 0 \implies \left\langle d_{ik}\left(\overline{s}_{ik}(\boldsymbol{\theta}) + B_i\right)^{-1}\right\rangle_k = \frac{T_{\text{exp}}}{K}. \tag{B.20}$$

Therefore if

$$B_i \gg \frac{K}{T_{\text{exp}}}\left\langle d_{ik}\right\rangle_k - \left\langle \overline{s}_{ik}(\boldsymbol{\theta})\right\rangle_k, \tag{B.21}$$

the conditional likelihood of B_i will be *far* smaller than the maximum (conditional) likelihood. The maximization approximation given by Equation (B.17) is not everywhere sufficiently accurate for centering the bounds of the integral domain, nor for the scaling term appearing in Equation (B.19). In particular, there may exist fringe regions of the prior hypervolume in which the signal from the target is exceptionally *bright* (e.g., small distances and intense surface radiation fields), meaning that application of approximation Equation (B.19) can result in *negative* background count rates, which is nonsensical, and if nonetheless applied can lead to total expected count-rate functions that are negative for some subdomain of the interval $[0,1] \ni \varphi$, meaning the likelihood function is undefined. We impose a hard physical prior boundary $\boldsymbol{B} = \boldsymbol{0}$ in order to impose meaning on the parameters $\boldsymbol{B}$ as defining a Poissonian background process;[9] it follows that a point calculated via application of Equation (B.19) that does not satisfy $\boldsymbol{B} \geq \boldsymbol{0}$ is not the point where the conditional likelihood function is maximal *within* the marginalization domain.

A critical and general remark to be made is that the conditional likelihood function is highly non-Gaussian on $\mathbb{R}$ *if the total expected count rate in any phase interval approaches zero as a function of the background count rate, where the likelihood rapidly decays and the curvature diverges.* If the expected total number of counts is zero when integrated over some subdomain of the interval $[0,1] \ni \varphi$ containing a finite number of observed events, the likelihood function *is* zero (provided that such information is not degenerated due to phase-binning) and thus not close to maximal. Such a scenario can occur if for instance the flux from the star falls to zero over some subinterval in phase, in which case background count rates approaching zero have associated likelihoods that decay rapidly to zero. If the phase-averaged signal from the star is bright, the background count rate that maximizes the

[9] In principle, discarding the meaning of the parameters $\boldsymbol{B}$ as Poissonian count rates, the maximum of the conditional likelihood function in a given channel could be permitted to exist at some negative $\boldsymbol{B}_i$, in which case the sum of the target count-rate function and $\boldsymbol{B}_i$ must by definition be greater than zero $\forall \varphi$.

conditional likelihood function may be small, but is strictly greater than zero.

Consider the expansion given by Equation (B.9): differentiating the conditional likelihood function along each background dimension we obtain

$$\frac{\partial\mathcal{L}}{\partial B_i}\approx\left.\frac{\partial\mathcal{L}}{\partial B_i}\right|_{y}+\frac{1}{2}\frac{\partial}{\partial B_i}\left[\left(B_i-B_i^{\prime}\right)^2\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{y}\right]=\left.\frac{\partial\mathcal{L}}{\partial B_i}\right|_{y}+\left(B_i-B_i^{\prime}\right)\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{y}. \tag{B.22}$$

Requiring an extremum (equivalent for L and $\mathcal{L}$ due to monotonicity of the logarithmic transformation),

$$\begin{aligned}0&\approx\left.\frac{\partial\mathcal{L}}{\partial B_i}\right|_{y}+\frac{1}{2}\frac{\partial}{\partial B_i}\left[\left(B_i-B_i^{\prime}\right)^2\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{y}\right]\\&=\left.\frac{\partial\mathcal{L}}{\partial B_i}\right|_{y}+\left(B_i-B_i^{\prime}\right)\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{y}\quad\Longrightarrow\quad B_i\approx B_i^{\prime}-\left.\frac{\partial\mathcal{L}}{\partial B_i}\right|_{y}\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{y}^{-1}.\end{aligned} \tag{B.23}$$

We therefore iteratively update B_i, given fiducial point $B_i^{\prime}$, evaluating the analytic derivatives at new points upon each iteration; the derivatives are given by Equation (B.12) and Equation (B.13). The curvature scale is thus used to approximate conditional likelihood function maximization, given the first-order derivative of the logarithm. We then need a termination criterion. Consider that if the conditional likelihood exhibits small departures from Gaussianity, the scale of the modal region of the function along each dimension is characterized almost entirely by the curvature. Suppose the conditional likelihood function in the i^{th} channel is approximated as

$$L(B_i;\boldsymbol{\theta})\propto\exp\left[-\frac{(B_i-\mu_i)^2}{2\sigma_i^2}\right]\quad\Longrightarrow\quad\sigma_i=\sqrt{2}\left[\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{B_i}\right]^{-\frac{1}{2}}. \tag{B.24}$$

We can therefore terminate the iterative procedure once $|\delta B_i|<\epsilon\sigma_i(B_i)$, for some small tolerance factor $\epsilon\ll 1$; for small departures from Gaussianity, this procedure converges extremely rapidly, in $\mathcal{O}(1)$ steps, and thus in a low-level language is entirely inexpensive. It follows that for termination:

$$|\delta B_i|:=\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{B_i}^{-1}\left|-\left.\frac{\partial\mathcal{L}}{\partial B_i}\right|_{B_i}\right|<\sqrt{2}\epsilon\left[\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{B_i}\right]^{-\frac{1}{2}}\quad\Longrightarrow\quad\left|-\left.\frac{\partial\mathcal{L}}{\partial B_i}\right|_{B_i}\right|<\sqrt{2}\epsilon\left[\left.\frac{\partial^2\mathcal{L}}{\partial B_i^2}\right|_{B_i}\right]^{\frac{1}{2}}. \tag{B.25}$$

For cases where the conditional likelihood function rapidly decays to zero for some background count rate, but the count rate that maximizes the function is in the near vicinity, one needs to be careful when performing such an iterative maximization procedure because the curvature and the first-derivative both diverge. As stated above, in general this is expected to only to occur for very bright model sources in fringes of the prior hypervolume, where the likelihood function is everywhere small relative to the global maximum. One therefore needs to protect against program instability, but need not be concerned too much with precisely calculating the marginal likelihood—the logarithm of the likelihood should be set to some large negative number if not inexpensively calculable.

Once we have a sufficiently accurate solution for the maximum likelihood point $\hat{B}_i$, we can choose bounds of the integral domain such as $(a_i, b_i) := \hat{B}_i \pm \alpha\sigma_i$, where $\alpha \sim 10$. Moreover, we can precompute the integrand scaling term

$$\max_{B_i}\left(-T_{\exp}B_i + \sum_j d_{ik}\ln c_{ik}\right) \approx -T_{\exp}\hat{B}_i + \sum_k d_{ik}\ln c_{ik}(\hat{B}_i), \tag{B.26}$$

such that for appropriate initial point B_i' and small ϵ, the exponent of the integrand is *at most* $\mathcal{O}(1)$. If the exponent is unexpectedly larger than this order, as a safeguard, (doubly) adaptive quadrature routines exist that can handle difficult integrands.

To obtain an initial value B_i' it is first necessary to check the lower bound on $\mathbb{R}$ at which the conditional likelihood function decays to zero. The expected count rate from the star is zero or positive in all phase intervals; the likelihood is defined and positive only for background count rates greater than the *negative* of the minimum of phase intervals of the count rate from the model star. If the count rate is precisely zero over some phase interval, the background count rate must be strictly positive; the initial value B_i' is therefore taken as the minimum over phase intervals of *finite* data counts, normalized by $T_{\exp}/n$, and multiplied by a small factor such as 10^{-2}.

Otherwise, applying Equation (B.19), if $B_i' < 0$, we impose that $B_i' = 0$, and if upon first iteration $\delta B_i < 0$, then it is known that $\hat{B}_i \equiv 0$, and σ_i is evaluated at zero; if on the other hand $\delta B_i > 0$, iteration proceeds from the improved initial point B_i'. If $\hat{B}_i - \alpha\sigma_i(\hat{B}_i) < 0$ we impose that $a_i := 0$; if on the other hand, the count rate from the model star is zero in a phase interval, $a_i := B_i'/10$, where B_i' is calculated from the minimum *finite* observed counts as described above.

Such a procedure rapidly generates very smooth numerical marginal likelihood functions over the entire prior hypervolume on the $\boldsymbol{\theta}$-subspace. The time to re-evaluate the function under a variation in the $\boldsymbol{B}$-subspace *at least* $\mathcal{O}(100)$ times faster than under a variation in the $\boldsymbol{\theta}$-subspace. In Figure B.1 we provide an example of how the background-marginalized likelihood behaves as a function of distance.

B.3 Mixed discrete-continuous model spaces

In this appendix we build general models and joint posterior distributions for Bayesian computation. Such models abstractly encompass instrument calibration uncertainty.

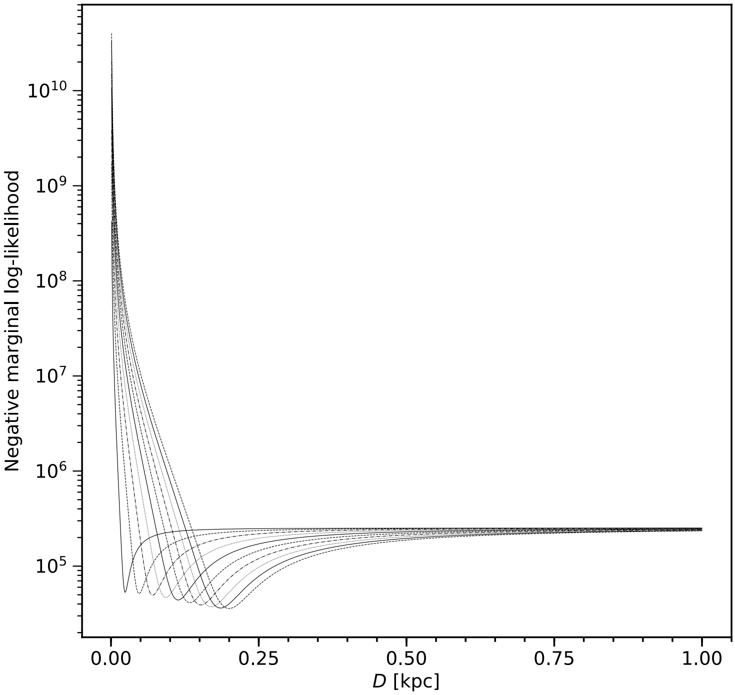

Figure B.1: An example of the *logarithm* of a background-marginalized likelihood function, conditional on a synthetic data set. The likelihood is displayed as a function of distance to a target source, for discrete fixed values of the angular radius of a circular hot spot on the surface. The parameter vector is otherwise identical to the injected vector. Note that the scale is double-logarithmic in likelihood. The marginal likelihood rapidly converges to a constant with increasing distance of the star, as required. For very small distances the logarithm of the marginal likelihood diverges to large negative numbers. Note that for an order of magnitude change in spot angular radius, the maximum (conditional marginal) likelihood distance changes by approximately an order of magnitude because the total flux from the star varies approximately linearly with the integrated proper area of the spot.

B.3.1 Model generalization

Let us define a discrete-continuous-mixed model $\mathscr{M}$ that is, in part,[10] a union $\bigcup_{m\in M}\mathcal{M}_m$, where $M\subset\mathbb{N}$ is the space of a discrete parameter m and where each $\mathcal{M}_m$ may be considered as a model within $\mathscr{M}$. We consider for readability a single discrete parameter, but the formalism in this appendix is extensible beyond a single discrete parameter. Let each model $\mathcal{M}_m$ define a continuous space $\Theta_m\subseteq\mathbb{R}^{n_m}$ such that $\mathscr{M}$ defines a space given by the union of Cartesian products

$$\Theta := \bigcup_{m\in M}\left(\{m\}\times\Theta_m\right). \tag{B.27}$$

The number of dimensions of the space Θ_m is given by n_m.

[10] A probability measure on the set M is also required for completion.

Let there exist a set of continuous target parameters $\boldsymbol{\theta}^\star \in \Theta^\star \subseteq \mathbb{R}^\star$ that are *shared* between the spaces Θ_m and of explicit interest for Bayesian computation. Further let us identify for each m the set of remaining parameters $\boldsymbol{\theta}_m$ that constitute Θ_m; a subset of the parameters θ_m may optionally be shared with one or more spaces $\Theta_{m'\neq m}$. Let us define the concept of parameter *sharing* more precisely by stating that the shared dimensions of Θ_m and $\Theta_{m'\neq m}$ are a single mathematical object; e.g., the $\boldsymbol{\theta}^\star$-dimensions form the space $\Theta^\star$. Each model $\mathcal{M}_m$ can define a distinct probability measure on the shared dimensions: for instance a joint prior probability density distribution $p(\boldsymbol{\theta}^\star \mid m)\colon M \times \Theta^\star \to \mathbb{R}$, where m is, in part, a discrete hyperparameter. Alternatively, the joint prior may be precisely the same probability measure $\forall m$, in which case $p(\boldsymbol{\theta}^\star)\colon \Theta^\star \to \mathbb{R}$.

Let us write the joint posterior distribution of the target parameters as

$$\begin{aligned} p(\boldsymbol{\theta}^\star \mid \boldsymbol{d}, \mathscr{M}, \mathscr{I}) &= \sum_{m\in M} \int p(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m \mid \boldsymbol{d}, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m \\ &= \sum_{m\in M} \int \frac{p(\boldsymbol{d} \mid \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m \mid \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \mid \mathscr{M}, \mathscr{I})} d\boldsymbol{\theta}_m, \end{aligned} \tag{B.28}$$

where the prior predictive probability of the data (the fully marginal likelihood) is given by

$$p(\boldsymbol{d} \mid \mathscr{M}, \mathscr{I}) = \sum_{m\in M} \int p(\boldsymbol{d} \mid \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m \mid \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m d\boldsymbol{\theta}^\star, \tag{B.29}$$

and the joint prior distribution is given by

$$\begin{aligned} p(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m \mid \mathscr{M}, \mathscr{I}) &= p(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_m \mid m, \mathscr{M}, \mathscr{I}) p(m \mid \mathscr{M}, \mathscr{I}) \\ &= p(\boldsymbol{\theta}^\star \mid \mathscr{M}, \mathscr{I}) p(\boldsymbol{\theta}_m \mid m, \mathscr{M}, \mathscr{I}) p(m \mid \mathscr{M}, \mathscr{I}), \end{aligned} \tag{B.30}$$

where $p(m \mid \mathscr{M}, \mathscr{I})$ is a probability *mass* distribution. Note that here we explicitly condition on both $\mathscr{M}$ and on prior information $\mathscr{I}$ (a union of independent observational data sets). The posterior is then written as

$$p(\boldsymbol{\theta}^\star \mid \boldsymbol{d}, \mathscr{M}, \mathscr{I}) = \frac{p(\boldsymbol{d} \mid \boldsymbol{\theta}^\star, \mathscr{M}, \mathscr{I}) p(\boldsymbol{\theta}^\star \mid \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \mid \mathscr{M}, \mathscr{I})}, \tag{B.31}$$

where

$$p(\boldsymbol{d} \mid \boldsymbol{\theta}^\star, \mathscr{M}, \mathscr{I}) = \underbrace{\sum_{m\in M} p(m \mid \mathscr{M}, \mathscr{I}) \int p(\boldsymbol{d} \mid \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}_m \mid m, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m}_{\text{marginal likelihood function}} \tag{B.32}$$

and letting $\mathcal{Z} := p(\boldsymbol{d} \,|\, \mathscr{M}, \mathscr{I})$,

$$\begin{aligned}\mathcal{Z} &= \int p(\boldsymbol{\theta}^\star \,|\, \mathscr{M}, \mathscr{I}) \sum_{m \in M} p(m \,|\, \mathscr{M}, \mathscr{I}) \int p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}_m \,|\, m, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m d\boldsymbol{\theta}^\star \\ &= \sum_{m \in M} p(m \,|\, \mathscr{M}, \mathscr{I}) \int p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}^\star \,|\, \mathscr{M}, \mathscr{I}) p(\boldsymbol{\theta}_m \,|\, m, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m d\boldsymbol{\theta}^\star \\ &= \sum_{m \in M} p(m \,|\, \mathscr{M}, \mathscr{I}) p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I}).\end{aligned} \tag{B.33}$$

Note that the posterior mass function on the discrete set M is then

$$p(m \,|\, \boldsymbol{d}, \mathscr{M}, \mathscr{I}) = \frac{p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I}) p(m \,|\, \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \,|\, \mathscr{M}, \mathscr{I})}. \tag{B.34}$$

Let us note the following limits: (i) if M has a solitary member m, then

$$p(\boldsymbol{\theta}^\star \,|\, \boldsymbol{d}, \mathscr{M}, \mathscr{I}) = \frac{p(\boldsymbol{\theta}^\star \,|\, \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \,|\, \mathscr{M}, \mathscr{I})} \int p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}_m \,|\, m, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m \tag{B.35}$$

where $p(\boldsymbol{d} \,|\, \mathscr{M}, \mathscr{I}) \equiv p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I})$; and (ii) if there do not exist any parameters $\boldsymbol{\theta}_m$, $\forall m$, such that $\Theta_m \equiv \Theta^\star$, then

$$p(\boldsymbol{\theta}^\star \,|\, \boldsymbol{d}, \mathscr{M}, \mathscr{I}) = \frac{p(\boldsymbol{\theta}^\star \,|\, \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \,|\, \mathscr{M}, \mathscr{I})} \sum_{m \in M} p(m \,|\, \mathscr{M}, \mathscr{I}) p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, m, \mathscr{M}), \tag{B.36}$$

where marginalization is not required to evaluate the likelihood function $p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, m, \mathscr{M})$, and

$$\begin{aligned}p(\boldsymbol{d} \,|\, \mathscr{M}, \mathscr{I}) &= \sum_{m \in M} p(m \,|\, \mathscr{M}, \mathscr{I}) \int p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, m, \mathscr{M}) p(\boldsymbol{\theta}^\star \,|\, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}^\star \\ &= \sum_{m \in M} p(m \,|\, \mathscr{M}, \mathscr{I}) p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I}).\end{aligned} \tag{B.37}$$

The primary sampling technique applied within *X-PSI* is that of nested sampling (see Section 3.4.1 and Section B.5). Therefore one requires a practical method for computation of the joint posterior distribution of the shared target parameters $\boldsymbol{\theta}^\star$. Typically, one proceeds by applying nested sampling to each member of a discrete set of models with associated continuous spaces, computing for each an evidence $\mathcal{Z}_m = p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I})$ and a set of weighted samples whose *weight* density approximates the joint posterior distribution

$$p(\boldsymbol{\theta}^\star \,|\, \boldsymbol{d}, i, \mathscr{M}, \mathscr{I}) = \frac{p(\boldsymbol{\theta}^\star \,|\, \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I})} \int p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}_m \,|\, m, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m. \tag{B.38}$$

Therefore, let us manipulate Equation (B.31) to obtain

$$p(\boldsymbol{\theta}^\star \,|\, \boldsymbol{d}, \mathscr{M}, \mathscr{I}) = \sum_{m\in M} \underbrace{\frac{p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I})p(m \,|\, \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \,|\, \mathscr{M}, \mathscr{I})}}_{p(m \,|\, \boldsymbol{d}, \mathscr{M}, \mathscr{I})} \underbrace{\frac{p(\boldsymbol{\theta}^\star \,|\, \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \,|\, m, \mathscr{M}, \mathscr{I})} \int p(\boldsymbol{d} \,|\, \boldsymbol{\theta}^\star, \boldsymbol{\theta}_m, m, \mathscr{M}) p(\boldsymbol{\theta}_m \,|\, m, \mathscr{M}, \mathscr{I}) d\boldsymbol{\theta}_m}_{\text{numerical weight-density approximation}}; \tag{B.39}$$

it follows that to marginalise over the discrete parameter m, one can *reweight* the nested samples with a posterior mass function over the set M:

$$p(m \,|\, \boldsymbol{d}, \mathscr{M}, \mathscr{I}) = \frac{p(m \,|\, \mathscr{M}, \mathscr{I})\mathcal{Z}_m}{\sum_{m\in M} p(m \,|\, \mathscr{M}, \mathscr{I})\mathcal{Z}_m}. \tag{B.40}$$

One does not therefore necessarily need to make a *decision* about which member $M_m \in \mathscr{M}$ to select in order to make inferential statements about shared parameters of interest, because the discrete nuisance parameter is marginalized. Indeed, if a subset of members $\mathcal{M}_m \in \mathscr{M}$ exhibit comparable posterior masses, that information manifests in the joint posterior distribution of the shared parameters.

For a set M of sufficiently small cardinality, applying nested sampling software to each member $\mathcal{M}_m \in \mathscr{M}$ is entirely feasible, and can be achieved in parallel in non-communicating processes given the appropriate computational resources. If, however, one is in the limit above—i.e., there are no *unshared* parameters—such that $\Theta_m \equiv \Theta^\star$, one could instead consider redefining $\mathscr{M}$ by connecting the discrete models $\mathcal{M}_m \in \mathscr{M}$ in a continuous manner, eliminating the need for a discrete parameter. The continuous parameters that connect the models are then jointly sampled (or analytically or numerically marginalized to reduce sampling-space dimensionality) with the target parameters $\boldsymbol{\theta}^\star$.

B.3.2 Discrete parametrization of an instrument

As a pertinent example let us suppose that a subset of nuisance parameters are defined to parameterize the operation of the instrument in response to incident radiation from point sources.

The matrix elements $\mathcal{R}^\star_{\ell ij}(\hat{r}_P)$ and $\mathcal{R}_{\ell ij}$ may *in principle* each be considered as (continuous) model parameters, and may be *measured*—i.e., statistically inferred given a data set acquired in some highly controlled environment such as a laboratory. The joint prior distribution of these parameters may in principle be highly informative, to the degree that it dominates any likelihood function given science observations, and the response matrix is thus *fixed* because considering the matrix elements as *variables* may significantly increase computational cost whilst minimally modulating the joint posterior distribution of the source parameters. Unless the instrument parameters can be analytically or numerically marginalized (perhaps with some judicious prior choices) to reduce the number sampling dimensions, the source contribution to the numerical likelihood function is in general too expensive for significant inflation of the number of parameter dimensions, as may be the case for a matrix with many more elements than source parameters.

If, on the other hand, the operation of the instrument is not highly constrained, one may be interested in performing joint estimation of source (and background) parameter and instrument parameters. For the reasons given above, however, it may be intractable to consider each response matrix element as a parameter; instead one might consider a lower-dimensional continuous parametrization, or a discrete or discrete-continuous-mixed parametrization.

As a toy example, let us consider a scenario in which there exist two fixed models for the *mean* instrument operation as appearing in Equation (3.24); let these models be assumed to bound ignorance to the true operation,[11] each of which specifies a response matrix $\mathcal{R}^{\star(m)}$ where $m \in \{1, 2\}$. Instead of identifying two models with the discrete parameter m, we could define a space of continuous parameters $\boldsymbol{\theta}$ such that the applied response matrix is given by $\mathcal{R}^{\star} := \mathcal{R}^{\star}(\boldsymbol{\theta}; \{\mathcal{R}^{\star(m)}\})$ where $m \in \{1, 2\}$ purely enumerates the matrices from which $\mathcal{R}^{\star}$ is built. Heeding the warning above regarding the dimensionality of the response matrix parametrization, let us consider a single-parameter family of response matrices to demonstrate what is effectively the simplest form of model for jointly estimating source and instrument properties. E.g., one could define

$$\mathcal{R}^{\star}(\boldsymbol{\theta}; \{\mathcal{R}^{\star(m)}\}) := w\mathcal{R}^{\star(1)} + (1 - w)\mathcal{R}^{\star(2)}, \tag{B.41}$$

where $w \in [0, 1]$; it is then the case that the element-wise mixture of the matrices is free, but the weights are uniform across all matrix elements. The joint posterior distribution of interest is then given by

$$p(\boldsymbol{\theta}^{\star} \mid \boldsymbol{d}, \mathscr{M}, \mathscr{I}) = \frac{p(\boldsymbol{\theta}^{\star} \mid \mathscr{M}, \mathscr{I})}{p(\boldsymbol{d} \mid \mathscr{M}, \mathscr{I})} \int p(\boldsymbol{d} \mid \boldsymbol{\theta}^{\star}, w, \mathscr{M}) p(w \mid \mathscr{M}, \mathscr{I}) dw, \tag{B.42}$$

where $p(\boldsymbol{d} \mid \mathscr{M}, \mathscr{I})$ is now the prior expectation of the likelihood function over the joint space of $(\boldsymbol{\theta}^{\star}, w)$. It follows that the marginalization integral in Equation (B.42) can be estimated with a *single* (nested) sampling process with respect to only continuous parameters $(\boldsymbol{\theta}^{\star}, w)$.

An added level of sophistication as demonstrated by Xu et al. (2014) is to consider a library of *calibration products*: a discrete set of functions of incident photon energy that collectively span the theoretical uncertainty in the channel-invariant effective area. Uncertainty in the redistribution matrix is beyond the scope of the work; the approach is thus consistent with writing the response matrix $\mathcal{R}^{\star}$ in terms of a discretized effective-area function and a redistribution matrix, and parametrising the effective-area function alone. The library is represented by a principal component analysis that summarizes the information within the

[11] Perhaps, e.g., there are two available response matrices that are discrepant in some manner, and a consensus cannot be reached as to which to condition on for source parameter estimation.

library in terms of a set of linearly independent objects; the compression of the number of library components can be used to construct instrument models with fewer parameters than there are source parameters, whilst maintaining the majority of the predictive complexity in the full library of calibration products.

B.4 Comments on priors and analysis workflow

An overarching notion that *X-PSI* is based upon is that pulse-profile modeling is a promising technique for deriving inferential statements about interior parameters—in particular those of cold dense matter EOS models—which may be assumed shared between members of an ensemble of NSs (e.g., Riley et al. 2018, and references therein; featured in Chapter 2, and hereafter R18). To this end, it is necessary to consider the problem of prior specification: an inherent component of any generative model. Moreover, prior specification is dependent on the parameter estimation workflow—i.e., on the sequence of intermediary computations which lead to inferential statements on parameters of interest. As discussed below and by R18, the workflow can take a number of forms involving exterior spacetime parameters (to which interior parameters are deterministically related via general relativistic gravity).

B.4.1 Overview

X-PSI offers a restricted set of models for a targeted star. The aim is to allow a user to couple low-level routines for signal computation to customizable high-level tools for building and applying a pulse-profile model, subject to tractability with open-source sampling software (see Section 3.4). A prior must be jointly specified for the parameters native to *X-PSI* and any ulterior model parameters, discrete or continuous, that a user defines.

Adopting viewpoints from Gelman et al. (2017), informative priors are specified without reference to a present or future likelihood function (i.e., without reference to observations to be analyzed), and encode all known information about a problem. In practice, however, it is mostly the case that complex data-generating processes are approximated in $\mathscr{M}$-complete or $\mathscr{M}$-open settings (e.g., Clarke et al. 2014), and it is far from obvious how to define an appropriate prior or even what a prior should purport to mean. Pragmatically, one may refer to a likelihood function (e.g., in order to construct parametrization-invariant definitions that are minimally informative from an information-theoretic perspective), or choices to impose inferential regularity and tractability may be made (Gelman et al. 2017).

Any application of *X-PSI* to data should be classifiable in either an $\mathscr{M}$-complete or an $\mathscr{M}$-open context: the target models supplied will at best withstand posterior predictive scrutiny when linked into a larger generative model, and stating the obvious, the user is warned that it may not be possible to build a model that is even satisfactory according to some predictive metric of performance. One reason for this is that target models in *X-PSI* are not fully self-consistent numerical simulations with parameterized boundary conditions and initial conditions, but are necessarily of a rudimentary phenomenological nature where physical intuition is lacking. The parameters native to *X-PSI* may generally be considered as bounded (or cyclic) mathematically. As an example, a substantial number of parameters might be defined as coordinates in a physical coordinate system and the underlying mathematical structure may guide prior definition. The inclination in spherical coordinates subtended by a distant observer to a star's rotation axis has compact support, and a flat prior may be intuited as natural under the notion of isotropy; however, the smaller the inclination (directions close to the rotation axis being smallest) the greater the degree of time-invariance of the distant signal on the rotational timescale, meaning pulsations may not be detected—a selection effect. We are therefore reminded that prior modeling is subsumed within generative modeling and all inferences are explicitly conditional on the set of models considered.

It is for the above reasons that, beyond noting that proper priors with compact support are usually more tractable for sampling applications, our purpose is not to advise on the details of priors. The single exception to the above rule is the focus of the remainder of this section.

B.4.2 Joint priors for exterior spacetime parameters

The exception concerns a workflow wherein interior parameters are not jointly sampled with any parameters that do not control spacetime structure, instead being sampled after all of these latter parameters have been marginalized over. Interior parameters as defined in R18 control local source matter properties and are, in the sense of locality, of a more fundamental nature than global exterior spacetime parameters that are unshared between NSs; in the astrophysical literature exterior parameters are typically directly constrained conditional on X-ray event data, with the intent to transform statistical information between exterior and interior parameter spaces. R18 treat this problem in detail, and to ensure rigor from a Bayesian perspective, advocate for a framework in which interior-parameter constraints are derived via direct posterior sampling conditional on (X-ray) data.

R18 also briefly consider the usage of information encoded by a set of samples from a marginal joint posterior of exterior parameters. It is suggested that if usage of some archival sample set is unavoidable—perhaps due to some form of resource limitation—then an approximate (marginal) likelihood function is to be constructed on some space. There appear to be a number of advantages inherent to such an approach, and it is also this workflow for which we can impose a formal condition on the form of the joint prior of exterior spacetime parameters.

First, R18 highlight the utility of information storage on an *intermediary* space of global exterior spacetime parameters associated with axisymmetric rotating NSs.[12] These parameters can be used to describe a NS without reference to an EOS. The interior source matter conditions map via integration to stable or unstable equilibrium spacetimes with known solution form, which in the exterior domain may be (Hartle 1967; Hartle & Thorne 1968): (i) decomposed into a countably infinite set of real numbers; and (ii) truncated at some order (in a small

[12] These parameters are defined such that they exist in the limit of rotating and non-rotating vacuum black hole solutions.

dimensionless rotation frequency) to define a small set of exterior parameters that ensure computational tractability of inference and with respect to which observations are expected to be informative. It follows that given such information, inferential statements may be derived regarding some general (discrete-continuous mixed) EOS model space because each model maps to an exterior spacetime with an approximated marginal likelihood function value (subject to conditions given below); in fact, the function need only be approximated up to an exterior-parameter invariant factor for the *relative* probabilities on the EOS model space to be derived.

Moreover, while a *subset* of interior parameters (and in a hierarchical context, hyperparameters) may be assumed to be *shared* between observed stars, the exterior spacetimes are *unshared* with respect to data generation. Therefore, if those exterior spacetimes are explicitly parameterized, the exterior parameters are unshared; further, a choice can be made to consider all nuisance parameters as unshared between the data-generating processes of distinct stars, whilst fixing any shared hyperparameters. A certain form of inferential parallelism then manifests: nuisance parameters may be marginalized over for each star separately (each by some research group), to draw a set of samples on a low-dimensional space of exterior parameters.

If shared interior parameters were to instead be jointly sampled with nuisance parameters (the *Interior Prior* paradigm in R18), it may become necessary to update posterior information sequentially, each time marginalising over unshared nuisance parameters, in order to both reduce the dimensionality of sampling spaces and organise workload with respect to research groups. In this case however, groups may be required to apply a mapping between interior and exterior parameters, which increases the complexity of each likelihood function evaluation; ideally a software implementation would be made available to ensure consistency whenever this shared mapping is applied.

The apparent advantages in terms of resource conservation and workload organisation should be clear even from this non-exhaustive discussion. We remark upon one other utility: not all research groups are interested in interior parameters—groups may instead be interested in implementing exterior-parameter posterior samples as an informative prior in a future analysis of observations from a given star. Moreover, if the (marginal) likelihood function is approximated and communicated, a group that implements that likelihood function can define the prior assumptions—the caveat being that doing so after gaining knowledge of the likelihood function structure requires appropriate justification (Gelman et al. 2017).

Despite such advantages, there are a number of conditions that, whilst mostly omitted and otherwise informally stated by R18, must be satisfied for such sample-based exterior parameter information to be usable for interior-parameter estimation. Ultimately, this approach to EOS inference may therefore be viewed as somewhat quixotic, and difficult to rigorously and generally formalize for practical application; below we merely continue commentary on the validity in practice of making inferential statements derived from exterior-parameter samples.

First, if the *joint* prior density distribution exhibits flat, compact support[13] on the some

[13] That is, within some closed hyperboundary, or disjoint set of closed hyperboundaries.

space, the joint posterior density distribution is proportional to the (marginal) likelihood function on that space. If the joint prior density, whilst proper,[14] is *not* flat on the sampling space, but exhibits flat, compact support on a space related to the sampling space via some invertible mapping, then samples may be transformed between spaces in order to approximate a (marginal) likelihood function via (weighted) posterior density estimation.[15] In summary, flatness is the strictly required property, and compact support is natural for exterior spacetime parameters based on general relativistic gravity. If, however, flatness is satisfied on the subspace of exterior parameters but properness is not jointly satisfied in the *full* parameter space, proof is required that the posterior is integrable; requiring prior integrability in the full space, however, eschews proof that the posterior is integrable for application of a sampler (amongst other benefits). Moreover, requiring compact support on the sampling space often improves sampling tractability. Given that flat proper priors are commonly invoked, requiring such conditions is far from outlandish; on the contrary the conditions may well be satisfied inadvertently. *Moreover, given that X-PSI may indeed be used to generate such exterior-parameter samples, the user may opt for a flat joint prior on the exterior-parameter space if it aligns with their research objectives.*

Even if flatness and integrability are satisifed, however, the inexorably finite degree of sampling noise means that the accuracy of such an approximation to the marginal likelihood function decays away from the *typical set* $\mathcal{T}_{\rm ext}$—i.e., away from the region of parameter space that, according to some probability density distribution, maps points within that region to some intermediate band in density and dominates expectation integrals (e.g., Betancourt 2017). If the set in which a draw is expected to lie is attributed $(1-\varepsilon)$ of the total probability mass, where $\varepsilon \ll 1$, the typical set is a uniquely defined as that wherein the density variation across contiguous regions is smallest; equipartition of the contained hypervolume thus results in some countable number of closed subsets with commensurate probability mass *relative* to other sets constructed so as to contain $(1-\varepsilon)$ of the total mass. The typical set does not coincide with maxima in the marginal likelihood function even when conditioned on a flat joint prior: the typical set is more generally disjoint and bounds maxima in parameter space, but the maxima are not members of $\mathcal{T}_{\rm ext}$. In other words, the typical set and the *highest density* set are not equivalent, with the most probable parameter vector being *atypical*.

On a theoretical level, a sampler is designed to draw (weighted) samples from the posterior distribution such that the asymptotic density of weights (in the limit of a countably infinite number of samples) everywhere converges to the posterior density. In practical applications this means efficiently generating samples from the typical set; strictly, in practical applications, the purpose of sampling is not to globally approximate the likelihood function. Having acknowledged this caveat, the likelihood function in the vicinity of the typical set—and more cautiously in the vicinity of the maxima—may be considered well-approximated up

[14] Integrable.

[15] Posterior density estimation may be numerical in nature, for instance via kernel density estimation, or may be achieved via some analytic approximation; for an effectively unimodal posterior one might invoke Gaussianity given an estimate of the mean vector and covariance matrix.

to a constant factor by the numerical density of weights. For a likelihood function with a single maximum, the expected Euclidean distance of a sample from a likelihood function maximum and the mean increases with dimensionality; for a given star the space of exterior parameters should be low-dimensional—i.e., after marginalization over a subspace of nuisance parameters. Nevertheless, inaccuracy may arise if, given some EOS model and the exact likelihood function, the typical set $\mathcal{T}_{\rm int}$ defined on a space of interior parameters does not map to (stable) exterior spacetime solutions in (or bounded by) the typical set $\mathcal{T}_{\rm ext}$, the region wherein the marginal likelihood function is accurately evaluable given the set of exterior-parameter samples.

If the joint prior distribution is known not to exhibit flat compact support, the marginal likelihood function is not proportional to the posterior density. In principle one should therefore attempt to determine to what degree the flatness is violated both in the vicinity of each maximum in the marginal likelihood function, and between those regions. If the (analytic) joint prior density is known one option is to approximate the marginal likelihood function (up to a constant factor) by taking the ratio of the marginal joint posterior density to the joint prior density function—although not without caveats (see section 2.3.4 of R18, and in particular the references therein).

Strictly *noninformative* priors, although rigorously—albeit quixotically—formalized from an information-theoretic perspective with respect to likelihood functions, are in our view here inadvisable due to tractability issues, both conceptual and computational (including integrability). Such priors seem rarely considered in applied fields (such as astrophysics) due to difficulty when working with complicated models. If we suppose a noninformative prior can be computed that, for instance, maximizes the mutual information between parameter space and data space (Berger et al. 2009), it certainly does not appear *guaranteed* that the degree of flatness for our purpose here is satisfactory. It may be the case, for instance, that although the density is approximately flat in the vicinity of each maximum in the marginal likelihood function, it is substantially different between those regions (e.g., consider a poorly specified, one-parameter model such that a likelihood function exhibits multiple maxima, and a reference prior, that of Jeffreys, such that the Fisher information differs to some degree between those maxima and their vicinities). Moreover, there is no consensus on the definition (nor philosophical meaning) of noninformative priors, and thus we have digressed in our pursuit of flatness and integrability.

Alternatively, one might consider the joint prior density as *weakly informative* with compact support (e.g., Gelman et al. 2017)—such a prior might also be referred as *vague* or *diffuse*. Such priors are often invoked with the aim of deriving a posterior predictive distribution that is locally insensitive to the details of the prior whilst restricting prior support to some compact set of conditional processes that do not generate synthetic data that appears strange in comparison to observed data (Gelman et al. 2017). To this end such priors seem to manifest as sufficiently flat relative to the likelihood function to permit learning from the data. Whilst we cannot comment universally on whether or not a prior considered as weakly informative can also be considered appropriate for sample-based marginal likelihood function approximation, we opine that such priors seem sufficiently practical and popular to warrant admission in the face of resource shortage. If the marginal joint prior distribution of exterior parameters is precisely known and considered weakly informative, we suggest an approximation given by the quotient of the (numerical) weight density and the prior density—i.e., up to a parameter-invariant factor, exchanging the weakly informative prior for a flat prior with equivalent *support*[16] (see also Neiswanger & Xing 2016; Abbott et al. 2019).

B.4.3 Further considerations for interior-parameter marginal likelihood function approximation

We now discuss interior-exterior spacetime solution matching criteria, assuming that a marginal likelihood function can be approximated given posterior samples on some space of exterior parameters. Any marginal likelihood function approximated on a space of exterior parameters must, strictly: (i) span the set of stable exterior solutions generated by the joint prior on the space of interior parameters, meaning that the compact support of the (flat) joint prior is a *superset* of those solutions; and (ii) be defined such that the conditional sampling distribution of the data depends on an exterior spacetime that is wholly defined by the *free* exterior parameters (multipole moments), meaning it does not include any fixed additional multipolar structure in the 2-surface cross-section nor in the metric.

An acceptable violation of the second condition immediately above would be the construction of a parameterized exterior spacetime using a quasi-universal relation between multipole moments. The purpose of such a construction is to reduce the dimensionality of the space of exterior solutions given the emergence of interior symmetries (Stein et al. 2014). For a given stellar rotation frequency, a model may be sensitive to inclusion or omission of higher-order exterior spacetime structure, but insensitive to the variation in that structure due to EOS variation. Quasi-universal relations are thus practically useful and are generally derived given a library of numerical spacetime solutions for some EOS family. In Section 3.2.3 we state that models within *X-PSI* necessarily embed an oblate radiating 2-surface within an ambient Schwarzschild spacetime. The user is free to implement a quasi-universal relation for the ellipticity (oblateness or polar radius) of the 2-surface; the default relation is that of AlGendy & Morsink (2014), written in terms of gravitational mass, coordinate equatorial radius, and coordinate rotation frequency.[17] *X-PSI* models are constructed given a *surface* radiation field; it follows that marginal likelihood functions on the joint space of *interior* parameters may be sensitive to the inclusion of lowest-order rotational deformation of the 2-surface, but insensitive to approximation of that rotational deformation via a quasi-universal relation.

An arguably more serious violation would occur if some subspace of free exterior parameters are marginalized over given a joint prior on that space; when evaluating the marginal likelihood function of a vector of interior parameters (with some associated stable interior spacetime solution), a single (truncated) exterior spacetime solution must be used that matches

[16] The subdomain of parameter space mapped to a non-zero density

[17] The coordinate rotation frequency may be fixed as discussed in R18.

to the interior solution. The subspace of model parameters marginalized over can therefore not include any exterior spacetime parameters that modify the structure of the spacetime.

Finally, we note that if interior parameters are jointly sampled with nuisance parameters (the *Interior Prior* paradigm in R18), then acceleration of the sequential updating process via group-by-group (marginal) likelihood function approximation given interior-parameter samples would encounter many of the problems discussed above, and the dimensionality of the space of shared interior parameters may be larger than that of the space of exterior parameters for each star.

B.5 Practical considerations for nested sampling

In this appendix we treat a number of technical matters pertaining to both prior density function implementation, and expectation integrals whose integrands are problematic for nested sampling. These details should be considered by the reader who intends to use *X-PSI*, but also may be of use to the reader who is generally interested in nested sampling theory and application.

B.5.1 Motivation

The original formulation of nested sampling (Skilling 2006) uniformly samples from the joint prior density distribution subject to an evolving likelihood function constraint. The MultiNest and PolyChord algorithms uniformly sample from a prior via an invertible mapping from a unit hypercube space to parameter space; whilst both algorithms utilize a unit hypercube to *inverse sample* by default, it is not necessary to strictly adhere to canonical prior inverse sampling rules if those rules are difficult for a given prior. It is, however, strictly necessary for the density of points within the unit hypercube to be uniform where the density is finite—i.e., within some disjoint set of likelihood function contours. Both algorithms implement cluster recognition algorithms to accelerate uniform sampling of the prior whilst imposing a likelihood function constraint, and these algorithms require the density of points within a cluster to be uniform. Aside from this condition, we are have some freedom to tinker with the way in which an invertible mapping between unit hypercube space and parameter space is constructed in order to simplify implementation of the prior, with the caveat that the efficiency and applicability of the integration algorithm may in some cases be adversely affected.

Canonically, when inverse sampling, an invertible map $\boldsymbol{x} \mapsto \boldsymbol{\theta}$ is constructed from the space of the unit hypercube $\mathcal{H}$ to the parameter space $\mathbb{R}^n$, such that a point $\boldsymbol{x} \in \mathcal{H}$ maps to a point $\boldsymbol{\theta} \in \mathbb{R}^n$ at which the joint prior density is *finite*. It is then the case that the unit hypervolume spanned by points $\boldsymbol{x} \in \mathcal{H}$ is equal to the total prior mass. Such a case is treated in detail in Feroz et al. (2009) and Handley et al. (2015). In this appendix we seek alternatives.

Suppose that $p(\boldsymbol{\theta})$ is a joint prior density distribution with some arbitrary form and compact support $\Theta \subset \mathbb{R}^n$; further let a hypersurface $\mathcal{B}$ be the boundary of Θ. The boundary

$\mathcal{B}$ may be trivial to evaluate but an analytic or numerical inverse-sampling map $\mathcal{H} \rightarrow \Theta$ may nevertheless be unwieldy. A pertinent example is any instance wherein $p(\boldsymbol{\theta})$ is of a numerical nature and, in order to minimize information loss and distortion, is to be used directly rather than approximately. If construction of a numerical map $\mathcal{H} \rightarrow \Theta$ as suggested by Feroz et al. (2009) proves difficult, one might opt for likelihood function redefinition. Such a prior $p(\boldsymbol{\theta})$ will generally be of an informative nature (e.g., Gelman et al. 2017), such as that arising when a (marginal) posterior density distribution is to be updated in a Bayesian sequential inference framework (e.g., R18). We address this problem in Section B.5.2.

Alternatively, let us suppose that this boundary is non-trivial to evaluate and can in principle consist of a set of disjoint hypersurfaces bounding disjoint regions of Θ, given by some set of constraint equations: typically it is inexpensive to determine if some $\boldsymbol{\theta} \in \mathbb{R}^n$ has finite local density (i.e., whether $\boldsymbol{\theta} \in \Theta$), but construction of an analytical inverse-sampling map $\mathcal{H} \rightarrow \Theta$ can be difficult, and one may prefer to circumvent implementation of a numerical map. We address this problem in Section B.5.3.

In practice both the form of $p(\boldsymbol{\theta})$ and the form of $\mathcal{B}$ might induce difficulties in construction of a map $\mathcal{H} \rightarrow \Theta$. The solution in Section B.5.3 is also a generalization appropriate for such a case.

Finally, in Section B.5.4 we give a detailed discussion on expectation integrals whose integrands are problematic for application of the canonical nested sampling algorithm. Problematic integrands may occur in practice when computing numerical likelihood functions: we offer workarounds for a particular class of such integrands, those that we encounter when implementing the default background model of *X-PSI* (see Section 3.2.10.3 and Section B.2). These workarounds do not require modification of nested sampler source code, and a completely rigorous treatment is beyond the scope of this work.

B.5.2 Likelihood function redefinition

As hinted by Feroz et al. (2009), we can subsume information encoded by the joint prior into a function whose expectation is to be calculated with respect to an alternative joint prior with (compact) support Θ. We now formalize this option.

Let us write $p(\boldsymbol{\theta}) = f(\boldsymbol{\theta})p^{\dagger}(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$, where the compact support of both the alternative joint density function $p^{\dagger}(\boldsymbol{\theta})$ and the dimensionless function $f(\boldsymbol{\theta})$ is also Θ. Let us further define $g(\boldsymbol{\theta}) \coloneqq L(\boldsymbol{\theta})f(\boldsymbol{\theta})$, such that the evidence is the expectation of the function $g(\boldsymbol{\theta})$ with respect to $p^{\dagger}(\boldsymbol{\theta})$:

$$\mathcal{Z} = \mathbb{E}_{p^{\dagger}(\boldsymbol{\theta})}[g(\boldsymbol{\theta})] = \int\limits_{\boldsymbol{\theta}\in\Theta} g(\boldsymbol{\theta})p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta} \coloneqq \int\limits_{a}^{b} g dY(g) = \int\limits_{0}^{1} g(Y)dY, \tag{B.43}$$

where

$$a = \min_{\boldsymbol{\theta}\in\Theta}[g(\boldsymbol{\theta})] \quad \text{and} \quad b = \max_{\boldsymbol{\theta}\in\Theta}[g(\boldsymbol{\theta})].$$

The aim is to construct an inverse-sampling map $\mathcal{H} \to \varTheta$, $\boldsymbol{x} \mapsto \boldsymbol{\theta}$ that is trivial, where $\mathcal{H}$ is the unit hypercube. We have the freedom to choose an analytic $p^{\dagger}(\boldsymbol{\theta})$ that approximates or captures information in $p(\boldsymbol{\theta})$, whilst circumventing the need for a numerical map. As highlighted by Feroz et al. (2009) for a *flat* $p^{\dagger}(\boldsymbol{\theta})$, a minor disadvantage of manipulating the definition of the likelihood function and prior as above is that for an *informative* prior $p(\boldsymbol{\theta})$, the information is subsumed within the redefined *expectation integrand* $g(\boldsymbol{\theta})$ that exhibits greater curvature, thus causing efficiency loss because compression of the prior mass Y defined with respect to $p^{\dagger}(\boldsymbol{\theta})$ requires a greater number of iterations. If flat, then $p^{\dagger}(\boldsymbol{\theta}) = 1/\mathcal{V}$ where

$$\mathcal{V} := \int\limits_{\boldsymbol{\theta}\in\varTheta} d\boldsymbol{\theta}. \tag{B.44}$$

Crucially, the numerical integration performed by nested samplers such as MultiNest natively generates quantities $w_i \sim Y_{i-1} - Y_i$ (given here in the crudest form; Skilling 2006) for a Riemann-sum approximation of $\mathcal{Z}$ as given by Equation (B.43), where $i \in \mathbb{N}$ enumerates iterations of the active-point replacement process. In this manipulated form, points $\boldsymbol{\theta}$ are drawn uniformly from $p^{\dagger}(\theta)$ within $\mathcal{B}$, subject to the constraint $g(\boldsymbol{\theta}) > g_i$ and given a uniform draw $\boldsymbol{x} \in \mathcal{H}$. The expected prior mass as a function of iteration number is then given by $\mathbb{E}[Y_i] = e^{-i/n}$ for n active points.

B.5.3 Inverse sampling with rejection for difficult support boundaries

Suppose that the bounding hypersurface $\mathcal{B}$ of the compact support $\varTheta$ of $p(\boldsymbol{\theta})$ is difficult to compute. Suppose then that the alternative joint density function $p^{\dagger}(\boldsymbol{\theta})$ is defined with compact support $\varTheta^{\dagger} \supseteq \varTheta$ whose boundary is the hypersurface $\mathcal{B}^{\dagger}$; let the construction of $p^{\dagger}(\boldsymbol{\theta})$ be such that an inverse-sampling map $\mathcal{H} \to \varTheta^{\dagger}$, $\boldsymbol{x} \mapsto \boldsymbol{\theta}$ is trivial to construct, where $\mathcal{H}$ is the unit hypercube. Let the joint density function $p(\boldsymbol{\theta})$ be defined on $\varTheta^{\dagger}$, and thus identically zero for $\boldsymbol{\theta} \in (\varTheta^{\dagger}\backslash\varTheta)$. Suppose we were to recast $\mathcal{Z}$ as an expectation integral with respect to $p^{\dagger}(\boldsymbol{\theta})$, via a function $f(\boldsymbol{\theta}) = p(\boldsymbol{\theta})/p^{\dagger}(\boldsymbol{\theta})$—i.e., $f : \varTheta^{\dagger} \mapsto \mathbb{R}$. The integrand g and integrator Y (the variable) would then be such that $Y \mapsto g$ is not everywhere well-defined on the interval $Y \in [0, 1]$, having a form illustrated in figure 14b of Skilling (2006), and the canonical nested sampling algorithm would not strictly be applicable (at least without modification). We discuss this problem in Section B.5.4.

Instead our aim is to inverse sample $p^{\dagger}(\boldsymbol{\theta})$ restricted to $\varTheta$: a point $\boldsymbol{x}$ is drawn uniformly from $\mathcal{H}$, operated on via $\boldsymbol{x} \mapsto \boldsymbol{\theta}$, and is discarded only if $\boldsymbol{\theta} \notin \varTheta$. Such a process may be considered as inverse sampling subject to a binary rejection condition given by the conjunction of the constraint equations for boundary $\mathcal{B}$. The points $\boldsymbol{x} \in \mathcal{H}$ such that $\boldsymbol{x} \mapsto \boldsymbol{\theta} \in \varTheta$ form subset of $\mathcal{H}$ if $\varTheta^{\dagger} \neq \varTheta$; the hypervolume $\mathscr{H}$ spanned by this subset depends on the properties of the map $\mathcal{H} \to \varTheta^{\dagger}$ and the properties of $\varTheta^{\dagger}\backslash\varTheta$. We require an extension of the definition of the function $f(\boldsymbol{\theta})$ for the case wherein $\varTheta^{\dagger} \supset \varTheta$: we write $f(\boldsymbol{\theta}) := \mathscr{H}p(\boldsymbol{\theta})/p^{\dagger}(\boldsymbol{\theta})$. The extended definition smoothly matches the definition given in Section B.5.2 for the case wherein

$\Theta^{\dagger} \equiv \Theta$—i.e., where $\mathscr{H} = 1$.

A practical case is for a flat $p(\boldsymbol{\theta})$ where $\mathcal{B}$ is known as a set of constraint equations that complicate construction of a map $\mathcal{H} \rightarrow \Theta$; we thus consider a flat $p^{\dagger}(\boldsymbol{\theta})$ and a hyper-rectangular $\mathcal{B}^{\dagger}$. The fractional hypervolume spanned by points $\boldsymbol{x} \in \mathcal{H}$ such that $\boldsymbol{x} \mapsto \boldsymbol{\theta} \in \Theta$ is given by the ratio, $\mathscr{H} = \mathcal{V}/\mathcal{V}^{\dagger}$, of the hypervolume of Θ to that of $\Theta^{\dagger}$.

X-PSI performs inverse sampling with rejection is as follows. The function that performs inverse sampling is a method bound to a `prior` object; the method itself is passed to MultiNest as a callback function. When MultiNest calls the callback, passing a point $\boldsymbol{x} \in \mathcal{H}$, the callback returns a point $\boldsymbol{\theta} \in \Theta^{\dagger}$ drawn uniformly from $p^{\dagger}(\boldsymbol{\theta})$. MultiNest handles the rejection operation indirectly via the native setting `logzero`. In *X-PSI*, the user must implement a built-in `__call__()` method for the `prior` object that evaluates a set of logical conditions to check whether or not $\boldsymbol{\theta} \in \Theta$. If $\boldsymbol{\theta} \in \Theta$, the `__call__()` method must return $\ln f(\boldsymbol{\theta}) - \ln \mathscr{H}$, evaluated simply as $\ln p(\boldsymbol{\theta}) - \ln p^{\dagger}(\boldsymbol{\theta})$;[18] if, on the other hand, $\boldsymbol{\theta} \notin \Theta$, the `__call__()` method must return negative infinity. In the latter case the point is to be rejected.

The `prior` callback is itself called by the `likelihood` callback; if point $\boldsymbol{x} \mapsto \boldsymbol{\theta}$ is to be rejected, the logarithm of the likelihood returned to MultiNest is less than `logzero` to force MultiNest to ignore it. If, on the other hand, $\boldsymbol{\theta} \in \Theta$, the `likelihood` callback proceeds to evaluate $\ln g(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) + \ln f(\boldsymbol{\theta})$, implicitly including the $\ln \mathscr{H}$ term.[19] We achieve this in an automated manner via straightforward Monte Carlo sample integration of the fractional hypervolume $\mathscr{H} \leq 1$ within $\mathcal{H}$ spanned by accepted points when inverse sampling $p^{\dagger}(\boldsymbol{\theta})$. The user need not be concerned with estimation of $\mathscr{H}$, but in principle the user is entirely free to replace the Monte Carlo sample integration routine with a more sophisticated integrator if considered necessary—perhaps if greater precomputation efficiency is desired. The default is to estimate $\mathscr{H}$; if however the user does not require $\mathscr{H}$—because, e.g., it is known to be unity—the user can deactivate estimation of $\mathscr{H}$, in which case $\mathscr{H}$ will default to unity.

Whilst not documented in the official MultiNest literature, it is briefly noted in the MultiNest `v2.13` changelog that points for which the logarithm of the likelihood function are less than `logzero` are treated as though they lie exterior to the support of the joint prior density function. Utilizing this behavior functions well in practice without error provided one compensates for potential efficiency loss as we detail below.

When a point is ignored, the ignorant MPI process waits for the next instruction to: (i) draw another point from the prior $p^{\dagger}(\boldsymbol{\theta})$ via inverse sampling; and (ii) subsequently determine

[18] Note that in the particular case that $\forall \boldsymbol{\theta} \in \Theta$, $p(\boldsymbol{\theta}) \equiv p^{\dagger}(\boldsymbol{\theta})/\mathscr{H}$ where $\mathscr{H} < 1$ (such as the practical case described above), then the user might require $\mathscr{H}$ in order to evaluate $p(\boldsymbol{\theta})$. For the example case, where $p(\boldsymbol{\theta})$ is flat, then $p(\boldsymbol{\theta}) = 1/\mathcal{V} = 1/(\mathscr{H}\mathcal{V}^{\dagger})$. A `try` clause with an `except` clause can be implemented in order to return: (i) $\ln f(\boldsymbol{\theta}) - \ln \mathscr{H} = -\ln \mathscr{H}$ if $\mathscr{H}$ has been estimated; or (ii) zero (the logarithm of a positive number, where the logarithm exceeds `logzero`) if, e.g., an `AttributeError` is raised because there does not yet exist a bound member of the `prior` object to which $\mathscr{H}$ is assigned once calculated. During the automated calculation of $\mathscr{H}$, calls to the `prior` object return zero when a point is accepted, and then $\mathscr{H}$ can be assigned to a member of the `prior` object to be accessed upon subsequent calls to the `prior` object in order to return $\ln f(\boldsymbol{\theta}) - \ln \mathscr{H} = -\ln \mathscr{H}$.

[19] If calls to the `prior` object return $\ln f(\boldsymbol{\theta}) - \ln \mathscr{H} = -\ln \mathscr{H}$, the `likelihood` callback returns $\ln g(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta})$ as required because the likelihood function was not redefined—i.e., $\ln f(\boldsymbol{\theta}) = 0$.

active-point candidacy via a likelihood evaluation if it is *not* rejected. At the commencement of sampling—the active-point generation phase—points are generated in this manner until the user-specified number of active points $\boldsymbol{\theta} \in \Theta$ is satisfied. The prior mass Y encompassed by this set of active points is taken to be unity at the outset of the active-point nested replacement phase, and the expected prior mass decays exponentially as a discrete function of the iteration (replacement) number. Evidence estimation is thus not adversely affected by this approach to inverse sampling.

Finally, we discuss a nuance to inverse sampling $p^{\dagger}(\boldsymbol{\theta})$ with rejection. The native sampling space remains as $\mathcal{H}$, in which points are uniformly distributed *within disjoint finite-density subdomains*, as required by the clustering and bound-construction algorithms. Moreover, these algorithms will be applied in $\mathcal{H}$ at the commencement of the replacement phase to determine if there are indeed disjoint regions with finite prior density that need to be bounded by disjoint hyper-ellipsoids to inverse-sample the prior efficiently. However, for the purpose of hyper-ellipsoidal decomposition of clusters in $\mathcal{H}$, the fractional hypervolume within $\mathcal{H}$ spanned by points $\boldsymbol{x} \mapsto \boldsymbol{\theta} \in \Theta$ is *not* taken to be $\mathscr{H}$, but instead *unity* at the outset, followed by exponential compression with iteration number. Whilst this assumption remains valid for evidence estimation, it can wreak havoc with sampling efficiency.

We are careful to handle the native efficiency setting, meaning there is effectively no additional computational cost to this type of prior implementation. We use the estimated fractional hypervolume $\mathscr{H}$ to modify the user-specified MultiNest sampling efficiency factor setting $\mathscr{E}$, and we then run MultiNest with the modified efficiency factor $\mathscr{E}/\mathscr{H}$. MultiNest uses the exponentially compressed reciprocal of the efficiency factor, $\mathscr{H}e^{-i/n_{\rm live}}/\mathscr{E}$, as the *lower bound* for the total hypervolume $\mathscr{H}_K$ in $\mathcal{H}$ bounded by the union of K hyper-ellipsoids at iteration i (whilst requiring the decomposition is *at least* minimally bounding)—see algorithm 1 of Feroz et al. (2009). The decomposition is computed by minimizing the total hypervolume $\mathscr{H}_K$ subject to $\mathscr{H}_K \geq \mathscr{H}e^{-i/n_{\rm live}}/\mathscr{E}$. If the user chooses $\mathscr{E} = 1$, we observe that, as required, $\mathscr{H}_K \geq \mathscr{H}e^{-i/n_{\rm live}}$. If we neglected such a modification to the user's specified efficiency factor, computing resource consumption for a given $\mathscr{E}$ would invariantly increase.

If $\mathscr{H}$ is excessively *small*, it becomes more computationally expensive to accurately estimate with small relative error, particularly in higher-dimensional unit hypercubes. Therefore we advocate that the boundary $\mathcal{B}^{\dagger}$ of the compact support $\Theta^{\dagger}$ be as close as possible to minimally bounding with respect to the compact support Θ as possible, whilst maintaining a simple form such as a hyper-rectangle; note however that this condition is not universally *sufficient*, but is sufficient for alternative density functions $p^{\dagger}(\boldsymbol{\theta})$ that are reasonably flat. The user has plenty of freedom to define an alternative density function as it is merely a computational tool without ulterior meaning. If $f(\boldsymbol{\theta}) = 1$, then the choice of $p^{\dagger}(\boldsymbol{\theta})$ is meaningful because it contributes to the definition of the prior density function $p(\boldsymbol{\theta})$. In this latter case weakly (or close to minimally) informative priors (refer to Section B.4 and the references therein) typically exhibit a reasonable degree of flatness; thus the condition regarding $\mathcal{B}$ bounding Θ should be sufficient.

Lastly, we note that the usual caveats regarding the accuracy of MultiNest's algorithm (for drawing uniformly from the prior subject to a likelihood constraint via a hyper-ellipsoidal decomposition) for the problem at hand (see Feroz et al. 2009) apply also in the context of inverse sampling the prior with rejection. The initial collection of active points approximately form a subset of $\mathcal{H}$ spanning hypervolume $\mathscr{H}$: the hyper-ellipsoidal decomposition needs to conform well[20] to this subset of $\mathcal{H}$ in order for sampling to be efficient and implementation-specific error low (Higson et al. 2019). The situation is congruent to that of the decomposition needing to conform well to a hypersurface through $\mathcal{H}$ on which the expectation integrand is invariant: MultiNest does not make a distinction between the distribution of active points in $\mathcal{H}$ having a certain form due to a likelihood constraint or due to the constraint that they report a likelihood value greater than `logzero`. Once the approximate minimum-hypervolume bounding ellipsoid spans a hypervolume less than unity, the clustering algorithm activates; if the hypervolume $\mathscr{H}$ is appreciably smaller than unity, MultiNest attempts to conform to the set of active points effectively immediately. Fortunately MultiNest is applicable to large class of problems, and given access to high-performance computing resources, we do not expect this to be a serious problem for *X-PSI*.

B.5.4 Problematic integrands for nested sampling

B.5.4.1 Overview

Problems arise for one-dimensional Riemann integration[21] of the expectation $\mathbb{E}_{p^{\dagger}(\boldsymbol{\theta})}[g(\boldsymbol{\theta})]$ if the map $Y \mapsto g$ is not everywhere well-defined on the interval $Y \in [0, 1]$, in particular when $g \mapsto Y$ is discontinuous as illustrated in figure 14b of Skilling (2006):[22] there exists a finite hypervolume within the compact support Θ of $p(\boldsymbol{\theta})$ over which the $g(\boldsymbol{\theta})$ is invariant.[23] The total prior mass associated with $g(\boldsymbol{\theta}) = g'$ is given from the Lebesgue perspective by the *finite* integrator element

$$dY(g') = \frac{1}{\mathscr{H}}\left[\int\limits_{g(\boldsymbol{\theta})\geq g'} p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta} - \int\limits_{g(\boldsymbol{\theta})>g'} p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta}\right] = \frac{1}{\mathscr{H}}\int\limits_{g(\boldsymbol{\theta})=g'} p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{B.46}$$

[20] Without a large fraction of the prior mass at a given iteration being spanned by two or more overlapping hyper-ellipsoids, which degrades sampling efficiency (Feroz et al. 2009).

[21] With respect to Y as given by Equation (B.43). Note that if instead a one-dimensional Riemann integral of the form

$$\mathcal{Z} = a + \int_{a}^{b} Y(g)dg, \text{ where } a = \min_{\boldsymbol{\theta}\in\Theta}[g(\boldsymbol{\theta})] \text{ and } b = \max_{\boldsymbol{\theta}\in\Theta}[g(\boldsymbol{\theta})], \tag{B.45}$$

where typically $a \ll b$, then $g \mapsto Y$ being discontinuous is not problematic. However, integration is then of a deterministic nature, where in order to evaluate $\mathcal{Z}$, one requires: (i) the limits of the integral domain; and (ii) the function $Y(g)$, an integral itself with domain $\Theta \subset \mathbb{R}^{n}$.

[22] Skilling (2006) highlighted such a map as "awkward" for nested sampling.

[23] As opposed to the subset of Θ over which $g(\boldsymbol{\theta})$ is invariant being, $\forall g$, *at most a hypersurface* (or disjoint set of such hypersurfaces) that by definition has zero hypervolume.

When formulated from the Lebesgue perspective, the integral remains well-defined and is evaluable: it is after all implicitly equivalent to straightforward Monte Carlo sample integration, which is perhaps more obviously well-defined.

In principle, if the function $g(Y)$ is *known* everywhere on the interval $Y \in [0, 1]$ and has compact range, it is not problematic that $g(\boldsymbol{\theta})$ function is invariant over some hypervolume in terms of posterior integrability: the evidence $\mathcal{Z}$ as given by the Riemann integral in Equation (B.43) would be well-defined even if $Y \mapsto g$ is continuous but information-losing. For instance, if the compact support of the function $g(\boldsymbol{\theta})$ is some subset of Θ there is simply no contribution to the integral from that subset.[24] However, if $g(Y)$ is non-analytic and Y cannot be everywhere explicitly and uniquely defined on the unit interval in relation to g, then the integrand is not well-defined everywhere on the integral domain, thus violating the ansatz that canonical nested sampling is founded upon: in other words, $g(Y)$ does not have connected support because it is undefined for at least one sub-interval of $Y \in [0, 1]$.

In order to make the Riemann integral well-defined, meaning must be given to Y on the sub-intervals of $Y \in [0, 1]$ wherein there does not exist a definition in terms of the prior mass subject to a constraint $g(\boldsymbol{\theta}) > g'$.[25] For the integral to be evaluable with a nested sampler one needs to extend the integrand $g(\boldsymbol{\theta})$ in a manner that approximates some arbitrary, numerically *close* function $g^{+}(\boldsymbol{\theta})$ in terms of which Y is everywhere well-defined on the unit interval, and thus whose expectation is Riemann-integrable via nested sampling.

B.5.4.2 Occurrence

In general, for a likelihood function $L(\boldsymbol{\theta})$ not modified by information from the prior density function $p(\boldsymbol{\theta})$, likelihood function invariance over a hypervolume may in principle suggest needless degeneracy within—or some other undesirable property of—the parametrization of the problem. Alternatively, if $L(\boldsymbol{\theta})$ *is* modified by information from a prior density function, thus defining the function $f(\boldsymbol{\theta}) = p(\boldsymbol{\theta})/p^{\dagger}(\boldsymbol{\theta})$,[26] the integrand function $g(\boldsymbol{\theta})$ may be identically zero for $\Theta^{\dagger}\backslash\Theta$—i.e., for a subset of the compact support $\Theta^{\dagger}$ that is here taken as the domain of the expectation integral.

More pragmatically, it may be the case that such behavior occurs or is deliberately induced when the likelihood function is of an approximative, numerical nature (Skilling 2006); in particular, if one can predict that the likelihood function is going to be very small for some $\boldsymbol{\theta}$ relative to the maxima in the function, one might choose to return an appropriately small likelihood value to avoid resource consumption whilst inducing negligible systematic error in the estimation of the evidence integral. As an example, we opt for such an approach when

[24] Note that if the compact support of the function $g(\boldsymbol{\theta})$ is not exactly the compact support $\boldsymbol{\Theta}$ of the intended density function $p(\boldsymbol{\theta})$—i.e., if the function $L(\boldsymbol{\theta})$ is identically zero for some subset of Θ—the nested sampling procedure interprets such behavior as *defining* the compact support of $p(\boldsymbol{\theta})$ via inverse sampling with rejection.

[25] One approach is to split the integral into a set of integrals whose domains are sub-intervals of $Y \in [0, 1]$ where Y is well-defined, apply nested sampling to each together with a correction to the prior mass for each; the sub-intervals where Y is undefined must then be treated separately. Such an approach would however be particularly cumbersome in all but special circumstances; see Section B.5.4.2 for an example.

[26] Note the deliberate omission of $\mathscr{H}$ here as distinct from in Section B.5.3.

implementing the default background model in *X-PSI* (see Section 3.2.10.3 and Section B.2): if the marginal likelihood is evidently going to be small, it is efficient to simply return an appropriately small value (e.g., slightly greater than `logzero`, where `logzero` is itself appropriately configured) in order to avoid wasting resources marginalising the likelihood function over the background parameters, which would otherwise be an expensive operation in some regions of parameter space despite the marginal likelihood function being relatively small.

Crucially, lest the prior density at points $\boldsymbol{\theta}$ be dependent on the data, the prior density function cannot be defined as a function of the likelihood function; this includes defining zero prior density *because* the likelihood function is *relatively* small. We cannot therefore *solely* apply the solution from Section B.5.3 in this case, which would be equivalent to defining zero prior density in regions over which the likelihood function is small, *after* gaining information from the data. The caveat here is that application of Section B.5.3 would necessitate a correction that reduces the evidence estimate (see Section B.5.4.4 for a very similar discussion on such corrections); given such a correction, the approach can be viewed as equivalent to splitting the Riemann integral into two parts, where one part contributes negligibly to $\mathcal{Z}$ and the other is an integral over a subset of $Y \in [0,1]$ on which $Y \mapsto g$ is diffeomorphic (at the level of numerical precision), and thus may be estimated via the canonical nested sampling algorithm. However, efficiencies will suffer if $\mathscr{H}$ is not accurately estimated, and if $\mathscr{H}$ depends on a numerical likelihood function its estimation can be computationally expensive. Clearly, if the evidence itself is not of interest, the importance weights are invariant to this detail and the solution from Section B.5.3 can be optionally applied.

B.5.4.3 The likelihood-extension scheme of Skilling (2006)

The decomposition of the prior density function into elements can only be statistically estimated. It cannot in general be known for (non-analytic) integrands $g(\boldsymbol{\theta})$ whether two points at which the integrand is equal exist on a hypersurface through $\mathbb{R}^n$ of constant and unique likelihood, or whether there does not exist a unique hypersurface. If numerical precision during sampling were infinite, no two active points within a finite set of samples drawn from the prior density distribution will exist within the same differential element of prior mass labelled uniquely with respect to the integrand. However, given that numerical precision is limited, it is not necessarily useful to concern oneself with the question of whether or not there exist, in the vicinity of the reported integrand value, truly unique hypersurfaces with respect to the integrand.

If two or more active points *do* report equivalent integrand values, Skilling (2006) offers, as a technical note, a general scheme in which equal-likelihood active points are extended with (pseudo)random labels of a *small* numeric nature (below the level of machine precision defined by the likelihood value itself); such an *extended* likelihood then allows the requisite ordering of points and nested sampling proceeds in the usual manner. Moreover, the ordering is not important for computation of $\mathcal{Z}$ because the likelihood effectively factors out, at the level of machine precision, of the associated terms in the Riemann summation.

We note however, that whilst MultiNest (effectively) employs the extended likelihood scheme of Skilling (2006) to resolve ties between *active* points in order to choose which to deactivate upon the i^{th} iteration, it does not natively extend the integrand in such a manner to resolve ties between draws from $p^{\dagger}(\boldsymbol{\theta})$ (with rejection if required) and the lower-bound at the i^{th} iteration, g_i (or L_i if $f(\boldsymbol{\theta}) := 1$). Consequently, the active-point *candidates*[27] are not generated in a manner that can handle an information-losing map $Y \mapsto g$—as illustrated in figure 14b of Skilling (2006)—which arises because the integrand is invariant numerically over a finite hypervolume in parameter space. We require a further extension to the function $g(\boldsymbol{\theta})$ to handle such a case.

The likelihood extension scheme of Skilling (2006) can also be applied here, but externally of MultiNest (a notion we will return to). Ideally, the extended likelihood labels are applied, below the level of machine precision, and stored in memory for all instances where: (i) a draw from the prior is generated and cannot be distinguished numerically from the lower-bound g_i; and (ii) if a point drawn from the prior becomes an active-point *candidate* without requiring the likelihood extension described in case (i), but ties in likelihood with an exisiting active point. In the former case the draw is extended in likelihood, and the deactivated point that defined the lower-bound g_i is also extended if not already labelled. In the latter case the likelihood of the candidate is extended, and so is the likelihood of the existing active point if it has not already been extended. These conditions ensure that all active points that tie in likelihood are extended with labels for ordering, and thus the active point to deactivate upon the i^{th} iteration (and thus define g_i) is proactively guarded against ambiguity.

Crucially, the numeric labels that extend the integrand at a given level of machine precision must be independent and identically distributed; for instance, one cannot assign a label to a draw (pending an active-point candidacy decision) based on the label assigned to the deactivated point that defined the lower-bound g_i. Labels applied in this manner impose an order necessary for the nested sampling algorithm to operate. To see why this is true, consider a large set of draws from the prior: those draws $\boldsymbol{\theta}$ for which $g(\boldsymbol{\theta}) = g_i$ are extended to $g^{+}(\boldsymbol{\theta}) = g_i + \epsilon(g_i)\ell$ where ℓ is a continuous random variable drawn from a sampling distribution $\ell \sim q(\ell)$ with compact support—such as the unit interval—such that the $\epsilon(g_i)\ell$ term cannot be resolved at the level of machine precision set by g_i.[28] The number of possible unique labels should always *far* exceed the number of points $\boldsymbol{\theta}$ at a given level $g(\boldsymbol{\theta})$ to which a label must be explicitly assigned and stored in memory. Now consider a *finite* prior mass element $dY(g_i)$ as given by Equation (B.46): a fraction y of draws from the prior at g_i are assigned labels greater than those assigned to a fraction $(1 - y)$ of all such draws, and thus a fraction y of the *mass* is assigned labels greater than those assigned to a fraction $(1 - y)$ of the mass. It follows that as the nested scheme progresses at $g(\boldsymbol{\theta}) = g_i$, the label of the extended lower-bound g_i^{+} lies above the labels of an increasing fraction $(1 - y)$ of the mass; thus the

[27] Where in general candidate generation workload is distributed amongst a set of MPI processes.

[28] The extension must be stored in memory via assignment to two variables: one variable, such as a `double` on a given architecture, for g_i, and another `double` for the label ℓ.

probability of a draw $\boldsymbol{\theta}$ from the prior reporting $g(\boldsymbol{\theta}) = g_i$ *and* not being outright rejected (and instead being listed as an active-point *candidate*[29]) is $ydY(g')$, which decays with iteration number.

Effectively *any* labelling function $\ell(\boldsymbol{\theta})$ can be applied—provided that an order is resultantly imposed—without adversely affecting the evidence integral. The sole condition to be adhered to is that the probability of a draw from the entire prior being used as the replacement *is equal to* the remaining prior mass at $g^{+}(\boldsymbol{\theta}) = g_i^{+}$;[30] above this is satisifed by identifying y as both the survival function of $q(\ell)$ and as the fraction of the prior mass at $g(\boldsymbol{\theta}) = g_i$ with larger labels. The expected evolution in prior mass Y is then given by

$$\mathbb{E}[Y_i] \sim \frac{e^{-i/n_{\text{live}}}}{\mathscr{H}} \int\limits_{g(\boldsymbol{\theta})\geq g_i} p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{B.47}$$

where i is the number of iterations *after* reaching g_i in the nested scheme. The nested sampler is “seeing” the following contribution to the evidence at $g(\boldsymbol{\theta}) = g_i$:

$$\Delta\mathcal{Z}(g_i) = -\frac{g_i}{\mathscr{H}} \int\limits_{\ell_a}^{\ell_b} \left[\int\limits_{\ell(\boldsymbol{\theta})>\ell'+d\ell'} p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta} - \int\limits_{\ell(\boldsymbol{\theta})>\ell'} p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta} \right], \tag{B.48}$$

where the variable is ℓ', and if, e.g., $\ell \sim \text{Uniform}(0, 1)$, then $(\ell_a, \ell_b) = (0, 1)$.[31]

The caveat here in practice is that MultiNest applies the lower-bound of g_i internally, at the level of precision set by g_i, and thus without extending the integrand. It thus becomes necessary to numerically extend the likelihood of a draw from the prior *above* the level of machine precision defined by the value of the lower-bound g_i, *if* MultiNest would otherwise reject the draw. We do not access the lower-bound g_i, which is defined and stored internally by MultiNest, and therefore we only suggest a solution for the case wherein the user understands where the integrand (which is defined externally of MultiNest) has been artificially induced to be numerically invariant over a finite hypervolume in parameter space, and in particular to be very small relative to maxima in the integrand (as reasoned above). The function $g(\boldsymbol{\theta})$ may well also be discontinuous at the sharp boundary of the region within which it is artificially set to some special small invariant value: such behavior would be similar to that illustrated in figure 14a of Skilling (2006), together with the plateau at the lower likelihood values as in figure 14b of Skilling (2006). In this case, the user can define a `likelihood` callback

[29] The probability of the draw being used as the replacement in the same iteration i depends on the number of MPI processes. The probability of the replacement draw, considering all MPI processes, reporting $g(\boldsymbol{\theta}) = g_i$ remains equal to $ydY(g')$.

[30] For more than one MPI process this probability naturally needs to be scaled according to the number of processes, but when considering all processes collectively the probability remains equal to the prior mass at $g^{+}(\boldsymbol{\theta}) = g_i^{+}$.

[31] Note that here ℓ is written as a function operating on $\mathbb{R}^n$: because the range of ℓ is one-dimensional (e.g., $\mathbb{R}$), the map $\mathbb{R}^n \to \mathbb{R}$ is information-losing, but in practice it is not necessary to distinguish between draws with equal ℓ values. This is analogous to not having to distinguish in practice between draws from a given *hypersurface* through $\mathbb{R}^n$ over which the integrand is invariant.

object that, if it would otherwise return the special value of $g(\boldsymbol{\theta})$, instead extends it with a label that can be resolved internally by MultiNest for ordering. For instance, if the special value is `logzero` (which will be some large negative number), labels could be defined such that $g^{+}(\boldsymbol{\theta}) = \texttt{logzero} \times (0.1 + 0.9\ell)$ where $\ell \sim \text{Uniform}(0, 1)$.

We also need to remark that a given implementation of nested sampling needs to draw uniformly from the prior according to $g(\boldsymbol{\theta}) \geq g_i$ for the above scheme to operate. However, implementations are generally designed to efficiently sample according to a hard likelihood constraint and resultantly *do not* sample uniformly *from* $\mathcal{H}$, but over disjoint subsets *of* $\mathcal{H}$. For the case discussed immediately above with MultiNest, the draws need to be subject to $g(\boldsymbol{\theta}) \geq \texttt{logzero}$. MultiNest imposes the condition that the hyper-ellipsoidal decomposition must be *bounding*; it follows that because the active points that report a random extended likelihood $g^{+}(\boldsymbol{\theta})$ near `logzero` are uniformly distributed over a subset of $\mathcal{H}$, the sampling region should not significantly contract until the region wherein $g(\boldsymbol{\theta}) = \texttt{logzero}$ (before extension) has largely been traversed. We reason that this must be true because the ℓ-ordering of the points is independent of $\boldsymbol{\theta} \in \Theta$ (and $\boldsymbol{x} \in \mathcal{H}$) and thus clustering is not expected to be recognized in $\mathcal{H}$; instead, at most a small number of hyper-ellipsoids should contiguously span both this region and the set of active points that report larger likelihoods. There may be repeated decomposition attempts because the hypervolume of the union of hyper-ellipsoids is greater than expected (see Feroz et al. 2009), slightly increasing computational expense until the region is traversed; however, performance should subsequently improve.

With the above scheme, PolyChord may well exit with an error that the integrand is non-deterministic: this is true, but is only problematic due to the nature and assumptions of the chordal sampling algorithm, which aims to improve sampling efficiency for well-behaved integrands. In summary, the viability of this form of likelihood extension is dependent on the specific algorithm implementing nested sampling (see Section B.5.5 for a pointer to an empirical demonstration that MultiNest can operate with such likelihood extension).

B.5.4.4 Corrections in the absence of a likelihood-extension scheme

Upon application of the default MultiNest configuration to a problematic map $Y \mapsto g$ as described in Section B.5.4, the statistical evolution of Y is erroneously treated. Above we have considered a pragmatic rationale for the induction of problematic behavior, but unless appropriate measures are taken, the performance of the nested sampler degrades—to be precise one explicitly introduces *systematic* error. We remark that this is in contrast to the effects highlighted in Section B.5 where without proper attention one can cause a degradation in the efficiency at which samples are drawn from the prior density function subject to an integrand constraint, but one does not *explicitly* suffer from remarkable systematic errors—in Section B.5 such error was of a more implicit nature, discussed elsewhere (in particular Feroz et al. 2009; Higson et al. 2018b; Higson 2018b; Higson et al. 2019).

Consider now a numerical function $g(\boldsymbol{\theta})$ that is invariant over some finite hypervolume within the support Θ of the prior density function. For simplicity let further us consider a case wherein: (i) the true value of $g(\boldsymbol{\theta})$ in this region is very small relative to maxima in $g(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$, and thus is artificially set to some exceedingly small default value such as MultiNest's `logzero` in order to avoid computational overhead; and (ii) the invariance occurs in the *fringes* of Θ near to where some weakly informative prior delimits (i.e., density falls to zero) what parameter vectors are considered "physical" (see also Section B.4).[32] In this case the first active point selected as a replacement—implicitly and internally by MultiNest via the likelihood extension of Skilling (2006) as described in Section B.5.4—will be a point $\boldsymbol{\theta}_i$ that reports $g(\boldsymbol{\theta}_i) = \texttt{logzero}$.

Suppose that we do not extend the function $g(\boldsymbol{\theta})$ as detailed in Section B.5.4. Now, crucially, when drawing from the prior to identify candidates for the replacement point, MultiNest will outright reject all draws that report $g(\boldsymbol{\theta}) = \texttt{logzero}$. The severe consequence is that if m of the initial n_{live} active points report $g(\boldsymbol{\theta}) = \texttt{logzero}$, MultiNest will commence by performing m iterations wherein the replacement point strictly report $g(\boldsymbol{\theta}) > \texttt{logzero}$. Once the first m points have been replaced, the n_{live} active points at the start of the $(m+1)^{th}$ iteration are drawn uniformly from the prior density function (with support Θ) subject to $g(\boldsymbol{\theta}) > \texttt{logzero}$, thus spanning the prior mass $Y(\texttt{logzero})$ given by

$$Y(\texttt{logzero}) = \frac{1}{\mathscr{H}} \int\limits_{g(\boldsymbol{\theta})>\texttt{logzero}} p^{\dagger}(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{B.49}$$

The canonical, expected evolution of Y is such that at the j^{th} iteration (Skilling 2006; Feroz et al. 2009; Handley et al. 2015)

$$\mathbb{E}[Y_j] = \prod_{i=1}^{j} \mathbb{E}[t_i] \approx e^{-j/n_{\text{live}}}, \tag{B.50}$$

in the absence of systematic implementation-specific nested sampling error (Higson et al. 2018b; Higson 2018b; Higson et al. 2019). The expected number of initial active points that report $g(\boldsymbol{\theta}) = \texttt{logzero}$ is given by

$$\mathbb{E}[m] = [1 - Y(\texttt{logzero})]\, n_{\text{live}}. \tag{B.51}$$

Once such points are replaced, the evolution of Y is such that

$$\mathbb{E}[Y_j] = \prod_{i=m}^{j} \mathbb{E}[t_i] \approx e^{-(j-m)/n_{\text{live}}} Y(\texttt{logzero}) = \underbrace{e^{-j/n_{\text{live}}}}_{\text{canonical factor}} \underbrace{e^{m/n_{\text{live}}} Y(\texttt{logzero})}_{\text{correction factor}} \tag{B.52}$$

where i here enumerates sampling iterations after replacement of the m initial active points that report $g(\boldsymbol{\theta}) = \texttt{logzero}$.

In principle $Y(\texttt{logzero})$ may be known effectively exactly via straightforward Monte

[32] The latter property is unnecessary but certainly plausible.

Carlo sample mean integration with respect to the prior of an integrand that implements the condition that $g(\boldsymbol{\theta}) > \texttt{logzero}$. However, determining whether the function $g(\boldsymbol{\theta})$ will be sufficiently small to neglect generally requires integrand evaluation, which for *X-PSI* models will generally be expensive. One can instead opt for a crude measure *given* the number m based on the sample files written to disk: with MultiNest the first m samples in the `ev.dat` file with the largest weights w_i will be empirically determinable. A rough estimate of can $Y(\texttt{logzero})$ can thus be obtained by inverting Equation (B.51) where n_{live} is user-specified. On the other hand, with an independent determination of $Y(\texttt{logzero})$ the correction factor will exhibit greater accuracy. Moreover, to improve efficiency, given an advance determination of $Y(\texttt{logzero})$ together with an *estimate* of m using Equation (B.51), one can modify (as in Section B.5.3) the efficiency setting of MultiNest by the *reciprocal* of the *correction factor* given in Equation (B.52).

To *reduce* systematic error in the estimate of $\mathcal{Z}$, one simply modifies the Riemann summation:

$$\begin{aligned}\ln\mathcal{Z} &\approx \text{logsumexp}\left(\{\ln w_i + \ln g_i\}_{i=1,\dots,j}\right)\\ &\to \frac{m}{n_{\text{live}}} + \ln Y(\texttt{logzero}) + \text{logsumexp}\left(\{\ln w_i + \ln g_i\}_{i=m+1,\dots,j}\right), \end{aligned} \tag{B.53}$$

where the first m terms in the summation: (i) are negligible when artificially using $g(\boldsymbol{\theta}) = \texttt{logzero}$; and (ii) should contribute negligibly to the evidence estimate if the *true* numerical values of $g(\boldsymbol{\theta})$ were used instead of `logzero`. Clearly the importance weights, being relative, are (almost exactly) invariant to such a modification.

B.5.4.5 Concluding comments on problematic integrands

If all points drawn uniformly with rejection from $p^{\dagger}(\boldsymbol{\theta})$ report identical values of the function $g(\boldsymbol{\theta})$, there are no points guiding the nested sampling process towards the subdomain(s) where $g(\boldsymbol{\theta})$ is higher (Skilling 2006), which is an issue particularly in high-dimensional spaces or if the required subdomain(s) occupy a small fraction of the prior support $\Theta^{\dagger}$. Indeed, MultiNest terminates if all active points report equal values of the integrand—see the v2.14 changelog; provided that there do exist active points that report higher integrand values, MultiNest proceeds.

On the other hand, if a likelihood extension scheme (or ordering or labeling) from Section B.5.4.3 is applied to *every* initial active point (because they report the same likelihood value) *before* it is communicated back to MultiNest, the MultiNest algorithm cannot be expected to operate properly to locate the region(s) where the integrand is deterministic and larger. It is therefore important desirable for any $\boldsymbol{\theta}$-invariance of the integrand to be associated with as small a prior mass as possible.

B.5.5 Concluding remarks

In the source code for *X-PSI* we supply a Jupyter notebook in which we apply MultiNest to compute an inexpensive evidence integral using a number of approaches. These calculations serve to empirically demonstrate the admittedly technical discussion in this appendix pertaining to nuances in how MultiNest in particular can be applied to expectation integrals. Implementation of the prior density function for nested sampling—and perhaps more rarely a numerical integrand with some artificially induced properties—can be a bane for certain problems, but we have shown above how the canonical configuration can be manipulated to conform to the problem at hand without any modification of the sampler source code.

Lastly we note that under likelihood function redefinition, the nested sampling process is organized such that the typical set (or posterior *bulk*; Feroz et al. 2009) is strictly integrated over whilst monotonically incrementing the modified likelihood function $g(\boldsymbol{\theta})$. It follows that if the function $f(\boldsymbol{\theta})$ is not flat on Θ, then the sample drawn with the maximum value of $g(\boldsymbol{\theta})$ is not in general the sample drawn with the maximum value of the likelihood function $L(\boldsymbol{\theta})$.

The target of computation in nested sampling is not some estimator of the parameter vector that maximizes the likelihood function; nevertheless, if the integration process is organized with respect to increasing $L(\boldsymbol{\theta})$, then in the limit that the remaining evidence (in the active points upon termination) $\Delta\mathcal{Z} \to 0$, the sample density (not weight density) is, loosely, maximal in the vicinity of the global maximum in the likelihood function assuming that peak is resolved in a given process. The relevant caveat here is that there is no restriction on the number of active points being invariant during nested sampling provided the statistical properties of sampling variables are monitored correctly, and indeed MultiNest reduces the number of active points when approaching termination (Feroz et al. 2009).[33]

If integration is performed whilst increasing $g(\boldsymbol{\theta})$ one therefore needs to modify the values of $g(\boldsymbol{\theta})$ associated with samples by the function $g(\boldsymbol{\theta})$ to recover the likelihood function values. If the function $g(\boldsymbol{\theta})$ is defined such that it is proportional to the *posterior* density function, the sample density in the limit that $\Delta\mathcal{Z} \to 0$ is, loosely, maximal in the vicinity of the global maximum in the posterior density function, and therefore the expected accuracy of the (already crude) maximum likelihood estimator given by the highest likelihood sample is degraded.

[33] The number of active points is also variant in *dynamic* nested sampling (Higson et al. 2018a).