Marginal likelihood computation for model selection and hypothesis testing: an extensive review

F. Llorente*, L. Martino**, D. Delgado*, J. Lopez-Santiago*

* Universidad Carlos III de Madrid, Leganés (Spain).

** Universidad Rey Juan Carlos, Fuenlabrada (Spain).

Abstract

This is an up-to-date introduction to, and overview of, marginal likelihood computation for model selection and hypothesis testing. Computing normalizing constants of probability models (or ratio of constants) is a fundamental issue in many applications in statistics, applied mathematics, signal processing and machine learning. This article provides a comprehensive study of the state-of-the-art of the topic. We highlight limitations, benefits, connections and differences among the different techniques. Problems and possible solutions with the use of improper priors are also described. Some of the most relevant methodologies are compared through theoretical comparisons and numerical experiments.

Keywords: Marginal likelihood, Bayesian evidence, numerical integration, model selection, hypothesis testing, quadrature rules, double-intractable posteriors, partition functions

1 Introduction

Marginal likelihood (a.k.a., Bayesian evidence) and Bayes factors are the core of the Bayesian theory for testing hypotheses and model selection [45, 74]. More generally, the computation of normalizing constants or ratios of normalizing constants has played an important role in statistical physics and numerical analysis [85]. In the Bayesian setting, the approximation of normalizing constants is also required in the study of the so-called double intractable posteriors [44].

Several methods have been proposed for approximating the marginal likelihood and normalizing constants in the last decades. Most of these techniques have been originally introduced in the field of statistical mechanics. Indeed, the marginal likelihood is the analogous of a central quantity in statistical physics known as the *partition function* which is also closely related to another important quantity often called *free-energy*. The relationship between statistical physics and Bayesian inference has been remarked in different works [2, 42].

The model selection problem has been also addressed from different points of view. Several criteria have been proposed to deal with the trade-off between the goodness-of-fit of the model and its simplicity. For instance, the Akaike information criterion (AIC) or the focused information criterion (FIC) are two examples of these approaches [76, 15]. The Bayesian-Schwarz information criterion (BIC) is related to the marginal likelihood approximation, as discussed in Section 2. The deviance information criterion (DIC) is a generalization of the AIC, which is often used in Bayesian inference [82, 83]. It is particularly useful for hierarchical models and it can be approximately computed when the outputs of a Markov Chain Monte Carlo (MCMC) algorithm are given. However, DIC is not directly related to the Bayesian evidence [71]. Another different approach, also based on information theory, is the so-called minimum description length principle (MDL) [36]. MDL was originally derived for data compression, and then was applied to model selection and hypothesis testing. Roughly speaking, MDL considers that the best explanation for a given set of data is provided by the *shortest description* of that data [36].

In the Bayesian framework, there are two main classes of sampling algorithms. The first one consists in approximating the marginal likelihood of different models. The second sampling approach extends the posterior space including a discrete indicator variable m, denoting the m-th model [8, 34]. Monte Carlo schemes working on this extended space (jumping in different spaces of parameters also with different dimensions) have been designed. We focus on the first approach.

In this work, we provide an extensive review of computational techniques for the marginal likelihood computation. The main contribution is to present jointly numerous computational schemes (introduced independently in the literature) with a detailed description under the same notation, highlighting their differences, relationships, limitations and strengths. It is also important to remark that parts of the presented material are also novel, i.e., no contained in previous works. We have widely studied, analyzed and jointly described, with a unique notation and classification, the methodologies presented in a vast literature from 1990s to the recent proposed algorithms (see Table 1). We also discuss issues and solutions when improper priors are employed. Therefore, this survey provides an ample covering of the literature, where we highlight important details and comparisons in order to facilitate the understanding of the interested readers and practitioners. The different techniques have been classified in four different families given below, and then described in details.

1.1 Problem statement and main notation

In many applications, the goal is to make inference about a variable of interest, $\mathbf{x} = x_{1:D_x} = [x_1, x_2, \dots, x_{D_x}] \in \mathcal{X} \subseteq \mathbb{R}^{D_x}$, where $x_d \in \mathbb{R}$ for all $d = 1, \dots, D_x$, given a set of observed measurements, $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$. In the Bayesian framework, one complete model $\mathcal{M}$ is formed by a likelihood function $\ell(\mathbf{y}|\mathbf{x}, \mathcal{M})$ and a prior probability density function (pdf) $g(\mathbf{x}|\mathcal{M})$. All the statistical information is summarized by the posterior pdf, i.e.,

$$\bar{\pi}(\mathbf{x}|\mathcal{M}) = p(\mathbf{x}|\mathbf{y}, \mathcal{M}) = \frac{\ell(\mathbf{y}|\mathbf{x}, \mathcal{M})g(\mathbf{x}|\mathcal{M})}{p(\mathbf{y}|\mathcal{M})}, \tag{1}$$

where

$$Z = p(\mathbf{y}|\mathcal{M}) = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x}, \mathcal{M})g(\mathbf{x}|\mathcal{M})d\mathbf{x}, \tag{2}$$

is the so-called marginal likelihood, a.k.a., Bayesian evidence. This quantity is important for model selection purpose, as we show below. However, usually $Z = p(\mathbf{y}|\mathcal{M})$ is unknown and difficult to approximate, so that in many cases we are only able to evaluate the unnormalized target function,

$$\pi(\mathbf{x}|\mathcal{M}) = \ell(\mathbf{y}|\mathbf{x},\mathcal{M})g(\mathbf{x}|\mathcal{M}). \tag{3}$$

Note that $\bar{\pi}(\mathbf{x}|\mathcal{M}) \propto \pi(\mathbf{x}|\mathcal{M})$ [45, 74]. For the sake of simplicity, hereafter we use the simplified notation $\bar{\pi}(\mathbf{x})$ and $\pi(\mathbf{x})$. Thus, note that

$$Z = \int_{\mathcal{X}} \pi(\mathbf{x})d\mathbf{x}. \tag{4}$$

Model Selection and testing hypotheses. Let us consider now M possible models (or hypotheses), $\mathcal{M}_1, ..., \mathcal{M}_M$, with prior probability mass $p_m = \mathbb{P}(\mathcal{M}_m)$, $m = 1, ..., M$. Note that, we can have variables of interest $\mathbf{x}^{(m)} = [x_1^{(m)}, x_2^{(m)}, \ldots, x_{D_m}^{(m)}] \in \mathcal{X}_m \in \mathbb{R}^{D_m}$, with possibly different dimensions in the different models. The posterior of the m-th model is given by

$$p(\mathcal{M}_m|\mathbf{y}) \quad = \quad \frac{p_m p(\mathbf{y}|\mathcal{M}_m)}{p(\mathbf{y})} \propto p_m Z_m \tag{5}$$

where $Z_m = p(\mathbf{y}|\mathcal{M}_m) = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x}_m, \mathcal{M}_m)g(\mathbf{x}_m|\mathcal{M}_m)d\mathbf{x}_m$, and $p(\mathbf{y}) = \sum_{m=1}^{M} p(\mathcal{M}_m)p(\mathbf{y}|\mathcal{M}_m)$. Moreover, the ratio of two marginal likelihoods $\frac{Z_m}{Z_{m'}}$

$$\frac{Z_m}{Z_{m'}} = \frac{p(\mathbf{y}|\mathcal{M}_m)}{p(\mathbf{y}|\mathcal{M}_{m'})} = \frac{p(\mathcal{M}_m|\mathbf{y})/p_m}{p(\mathcal{M}_{m'}|\mathbf{y})/p_{m'}}, \tag{6}$$

also known as *Bayes factors*, represents the posterior to prior odds of models m and m'. If some quantity of interest is common to all models, the posterior of this quantity can be studied via *model averaging* [38], i.e., a the complete posterior distribution as a mixture of M partial posteriors linearly combined with weights proportionally to $p(\mathcal{M}_m|\mathbf{y})$ (see, e..g, [61, 87]). Therefore, in all these scenarios, we need the computation of Z_m for all $m = 1, ..., M$. In this work, we describe different computational techniques for calculating Z_m, mostly based on Markov Chain Monte Carlo (MCMC) and Importance Sampling (IS) algorithms [74]. Hereafter, we assume proper prior $g(\mathbf{x}|\mathcal{M}_m)$. Regarding the use of *improper priors* see Section 6. Moreover, we usually denote Z, $\mathcal{X}$, $\mathcal{M}$, omitting the subindex m, to simplify notation. It is important also to remark that, in some cases, it is also necessary to approximate normalizing constants (that are also functions of the parameters) in each iteration of an MCMC algorithm, in order to allow the study of the posterior density. For instance, this is the case of the so-called double intractable posteriors [44].

Reversible jump approach. Other sampling approaches include a discrete model indicator variable m which denotes the m-th model, i.e., considering an extended posterior space [8, 34]. An MCMC method working on the extended space is then designed and run. In the well-known reversible jump MCMC [34], a Markov chain is generated allowing jumps between models with parameter spaces with possibly different dimensions. However, generally, these methods are difficult to tune and the mixing of the chain can be poor [37]. For further details, see also the interesting works [18, 33, 16]. In this work, we focus on the direct marginal likelihood approximation.

1.2 A general overview

After a depth revision of the literature, we have recognized four main families of techniques, described below. We list them in order of complexity, from the simplest to the most complex underlying main idea. However, each class can contain both simple and very sophisticated algorithms.

Family 1: *Deterministic approximations***.** These methods consider an analytical approximation of the function $\bar{\pi}(\mathbf{x})$. The Laplace method and the Bayesian Information Criterion (BIC), belongs to this family.

Family 2: *Methods based on density estimation***.** This class of algorithms uses the equality

$$\widehat{Z} = \frac{\pi(\mathbf{x}^*)}{\widehat{\bar{\pi}}(\mathbf{x}^*)}, \tag{7}$$

where $\widehat{\bar{\pi}}(\mathbf{x}^*) \approx \bar{\pi}(\mathbf{x}^*)$ represents an estimation of the density $\bar{\pi}(\mathbf{x})$ at some point $\mathbf{x}^*$. Generally, the point $\mathbf{x}^*$ is chosen in a high-probability region. The techniques in this family differ for the procedure employed for obtaining the estimation $\widehat{\bar{\pi}}(\mathbf{x}^*)$. Some famous example is the Chib's method [11].

Family 3: *Importance sampling (IS) schemes.* The IS methods are based on rewriting Eq. (2) as an expected value w.r.t. a simpler normalized density $\bar{q}(\mathbf{x})$, i.e., $Z = \int_{\mathcal{X}} \pi(\mathbf{x}) d\mathbf{x} = E_{\bar{q}}\left[\frac{\pi(\mathbf{x})}{\bar{q}(\mathbf{x})}\right]$. This is the most considered class of methods in the literature, containing numerous variants, extensions and generalizations. We devote Sections 3-4 to this family of techniques.

Family 4: *Methods based on a vertical representation.* These schemes rely on changing the expression of $Z = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x}) g(\mathbf{x}) d\mathbf{x}$ (that is a multidimensional integral) to equivalent unidimensional integrals [70, 89, 80]. Then, a quadrature scheme is applied to approximate this unidimensional integral. The most famous example is the nested sampling algorithm [80]. Section 5 is devoted to this class of methods.

1.3 Other reviews and software packages

The related literature is rather vast. In this section, we provide a brief summary that intends to be illustrative rather than exhaustive, by means of Table 1. The most relevant (in our opinion) and related surveys are compared according to the topics, material and schemes described in the work. The proportion of covering and overlapping with this work is roughly classified as "partial" ◇, "complete" $\surd$, "remarkable" or "more exhaustive" work with ★. From Table 1, we can also notice that completeness of this work. We take into account also the completeness and the depth of details provided in the different derivations. The Christian Robert's blog deserves a special mention (`https://xianblog.wordpress.com`), since Professor C. Robert has devoted several entries of his blog with very interesting comments regarding the marginal likelihood estimations and related topics.

Software packages. Currently, there is specific software aimed at performing Bayesian model choice, and more specifically, the computation of marginal likelihoods, for general models. Different R packages are available. We can find the *bridgesampling* package [35], which implements the bridge sampling estimator (see Sect. 3.3) for computing marginal likelihoods given only a posterior sample and the log posterior function. More generally, under CRAN Task View: Bayesian Inference, there are several packages for performing Bayesian inference that also include functions to estimate the marginal likelihood, for instaince, *MCMCpack* [48] and *LaplacesDemon* [84]. *MCMCpack* allows for fitting a different number of models and calculate the corresponding marginal likelihoods mainly by Laplace approximation and Chib's method (see Section 2). The *LaplacesDemon*'s main function produces an estimate of the marginal likelihood using the harmonic mean, generalized harmonic mean, Laplace Metropolis or importance sampling (see Sections 2 and 3). Additionolly, the review by [26] provides with the codes in R used to implement the different methods discussed in that review.

Structure of the paper. Section 2 is devoted to describe methods belonging to family 1 and family 2. Section 3 and Section 4 introduce the methods in family 3. More specifically, in Section 3 we consider IS approaches using one, two or more proposal densities. Some of them require also the use of MCMC techniques. In Section 4, we describe more sophisticated methods that combine the IS schemes with the MCMC algorithms. The vertical approach corresponding to Family 4 above is described in Section 5. In Section 6, we present how to deal with hypotheses testing and model selection problems when the employed prior densities are improper. Section 7 contains some theoretical example and numerical experiments. In Section 8, we conclude with a final summary and discussion.

2 Methods based on deterministic approximations and density estimation

In this section, we consider approximations of $\bar{\pi}(\mathbf{x})$, or its unnormalized version $\pi(\mathbf{x})$, in order to obtain an estimation Z. In a first approach, the methods consider $\bar{\pi}(\mathbf{x})$ or $\pi(\mathbf{x})$ as a function, and try to obtain a good approximation given another parametric or non-parametric family of functions. Another approach consists in approximating $\bar{\pi}(\mathbf{x})$ only in one specific point $\mathbf{x}^*$, i.e., $\widehat{\bar{\pi}}(\mathbf{x}^*) \approx \bar{\pi}(\mathbf{x}^*)$ ($\mathbf{x}^*$ is usually chosen in high posterior probability regions), and then using the identity,

$$\widehat{Z} = \frac{\pi(\mathbf{x}^*)}{\widehat{\bar{\pi}}(\mathbf{x}^*)}. \tag{8}$$

The latter scheme is often called *candidate's estimation.*

Laplace's method. Let us define $\widehat{\mathbf{x}}_{\text{MAP}} \approx \mathbf{x}_{\text{MAP}} = \arg\max \bar{\pi}(\mathbf{x})$ and consider a Gaussian approximation of $\bar{\pi}(\mathbf{x})$ around $\widehat{\mathbf{x}}_{\text{MAP}}$, i.e.,

$$\widehat{\bar{\pi}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\widehat{\mathbf{x}}_{\text{MAP}}, \widehat{\mathbf{\Sigma}}), \tag{9}$$

Table 1: Covering of the considered topics of other surveys or works (◇: partial, $\surd$: complete, ★: remarkable or more exhaustive). We take into account also the completeness and the depth of details provided in the different derivations. To be more precise, in the case of Section 4.1, we have also considered the subsections.

Surveys	Sect. 2	Sect. 3			Sect. 4						Sect. 5			Sect. 6
		3.1	3.2	3.3	4.1				4.2	4.3	5.1	5.2	5.3	
[27, 39]	✓	◇												
[31, Ch. 10]	◇	✓	◇											
[63]		◇	★	◇										
[19]	★	✓	✓											
[9][10, Ch. 5]	◇	◇	✓	✓										
[28]		✓	★	★										
[4]	✓	✓												
[88]		◇		◇		◇								
[47]	◇	✓	◇											
[75]		✓	✓									✓		
[26]	◇	◇		◇		✓						✓		
[1]	◇	◇	✓											
[70]		◇		◇							✓	★	★	
[77]	✓	✓										◇		
[40]	◇	◇		✓		✓						✓		
[46]	◇	◇		✓								★		
[90]	★	✓	✓											
[49]								◇	★					
[5, 6]		◇								★				
[72]		◇		✓		◇	◇					◇		
[69, 3]														★

with $\widehat{\mathbf{\Sigma}} = (-D_x^2 \log \pi(\widehat{\mathbf{x}}_{\text{MAP}}))^{-1}$, which is the negative inverse Hessian matrix at $\widehat{\mathbf{x}}_{\text{MAP}}$. Replacing in Eq. (8), with $\mathbf{x}^* = \widehat{\mathbf{x}}_{\text{MAP}}$, we obtain the Laplace approximation

$$\widehat{Z} = \frac{\pi(\widehat{\mathbf{x}}_{\text{MAP}})}{\mathcal{N}(\widehat{\mathbf{x}}_{\text{MAP}}|\widehat{\mathbf{x}}_{\text{MAP}}, \widehat{\mathbf{\Sigma}})} = (2\pi)^{\frac{D_x}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}} \pi(\widehat{\mathbf{x}}_{\text{MAP}}). \tag{10}$$

This is equivalent to the classical derivation of Laplace's estimator, which is based on expanding the $\log \pi(\mathbf{x}) = \log(\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}))$ as quadratic around $\widehat{\mathbf{x}}_{\text{MAP}}$ and substituting in $Z = \int \pi(\mathbf{x})d\mathbf{x}$, that is,

$$\begin{aligned} Z = \int \pi(\mathbf{x})d\mathbf{x} &= \int \exp\{\log \pi(\mathbf{x})\} d\mathbf{x} && (11) \\ &\approx \int \exp\left\{\log \pi(\widehat{\mathbf{x}}_{\text{MAP}}) - \frac{1}{2}(\mathbf{x} - \widehat{\mathbf{x}}_{\text{MAP}})^T \widehat{\mathbf{\Sigma}}^{-1}(\mathbf{x} - \widehat{\mathbf{x}}_{\text{MAP}})\right\} d\mathbf{x} && (12) \\ &= (2\pi)^{\frac{D_x}{2}} |\widehat{\mathbf{\Sigma}}|^{\frac{1}{2}} \pi(\widehat{\mathbf{x}}_{\text{MAP}}). && (13) \end{aligned}$$

In [43], it is suggested using the posterior simulation output to estimate the quantities $\widehat{\mathbf{x}}_{\text{MAP}}$ and $\widehat{\mathbf{\Sigma}}$ using samples generated by a Metropolis-Hastings algorithms. The resulting method is called "Laplace-Metropolis" estimator. The authors in [19] present different variants of the Laplace's estimator.

Bayesian-Schwarz information criterion (BIC). Let us define $\widehat{\mathbf{x}}_{\text{MLE}} \approx \mathbf{x}_{\text{MLE}} = \arg\max \ell(\mathbf{y}|\mathbf{x})$. The following quantity

$$\text{BIC} = D_x \log D_y - 2 \log \ell(\mathbf{y}|\widehat{\mathbf{x}}_{\text{MLE}}), \tag{14}$$

was introduced by Gideon E. Schwarz in [79], where D_x represents the number of parameters of the model ($\mathbf{x} \in \mathbb{R}^{D_x}$), D_y is the number of data, and $\ell(\mathbf{y}|\widehat{\mathbf{x}}_{\text{MLE}})$ is the estimated maximum value of the likelihood function. The value of $\widehat{\mathbf{x}}_{\text{MLE}}$ can be obtained using an MCMC scheme. The BIC expression can be derived similarly to the Laplace's method, but this time with a second-order Taylor expansion of the log likelihood around its maximum $\mathbf{x}_{\text{MLE}}$ and considering uniform improper priors over the parameters, resulting in

$$Z \approx \widehat{Z} = \exp\left(\log \ell(\mathbf{y}|\widehat{\mathbf{x}}_{\text{MLE}}) - \frac{D_x}{2} \log D_y\right) = \exp\left(-\frac{1}{2}\text{BIC}\right), \tag{15}$$

and BIC $\approx -2\log Z$. Then, smaller BIC values are associated to better models. Note that BIC clearly takes into account the complexity of the model since higher BIC values are given to models with more number of parameters D_x. Namely the penalty $D_x \log D_y$ discourages overfitting, since increasing the number of parameters virtually always improves the goodness of the fit. Other criteria can be found in the literature, such as the well-known Akaike information criterion (AIC),

$$\text{AIC} = 2D_x - 2 \log \ell(\mathbf{y}|\widehat{\mathbf{x}}_{\text{MLE}}).$$

However, they are not an approximation of the marginal likelihood Z and they are usually founded on information theory derivations. Generally, they have the form of $c_p - 2\log \ell(\mathbf{y}|\widehat{\mathbf{x}}_{\text{MLE}})$ where the penalty term c_p of the model complexity changes in each different criterion (e.g., $c_p = D_x \log D_y$ in BIC and $c_p = 2D_x$ in AIC). Another example that uses MCMC samples is the Deviance Information Criterion (DIC), i.e.,

$$\text{DIC} = -\frac{4}{N}\sum_{i=1}^{N} \log \ell(\mathbf{y}|\mathbf{x}_n) - 2\log \ell(\mathbf{y}|\bar{\mathbf{x}}), \qquad \text{where} \quad \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_n, \tag{16}$$

and $\{\mathbf{x}_n\}_{n=1}^N$ are outputs of an MCMC algorithm [82]. In this case, note that $c_p = -\frac{4}{N}\sum_{i=1}^{N} \log \ell(\mathbf{y}|\mathbf{x}_n)$. DIC is considered more approximate for hierarchical models than AIC, BIC [82] (but is not directly related to the marginal likelihood [71]).

Kernel density estimation (KDE). KDE can be used to approximate the value of the posterior density at a given point $\mathbf{x}^*$, and then consider $Z \approx \frac{\pi(\mathbf{x}^*)}{\bar{\pi}(\mathbf{x}^*)}$. For instance, we can build a kernel density estimate (KDE) of $\bar{\pi}(\mathbf{x})$ based on M samples distributed according to the posterior (obtained via an MCMC algorithm, for instance) by using M normalized kernel functions $k(\mathbf{x}|\boldsymbol{\mu}_m, h)$ (with $\int_{\mathcal{X}} k(\mathbf{x}|\boldsymbol{\mu}_m, h)d\mathbf{x} = 1$ for all m) where $\boldsymbol{\mu}_m$ is a location parameter and h is a scale parameter,

$$\widehat{\pi}(\mathbf{x}^*) = \frac{1}{M}\sum_{m=1}^{M} k(\mathbf{x}^*|\boldsymbol{\mu}_m, h), \quad \{\boldsymbol{\mu}_i\}_{m=1}^{M} \sim \bar{\pi}(\mathbf{x}) \quad \text{(e.g., via MCMC)}. \tag{17}$$

The estimator is $\widehat{Z} = \frac{\pi(\mathbf{x}^*)}{\widehat{\pi}(\mathbf{x}^*)}$ where the point $\mathbf{x}^*$ could be $\widehat{\mathbf{x}}_{\text{MAP}}$. If we consider N different points $\mathbf{x}_1, ..., \mathbf{x}_N$ (selected without any specific rule) we can also write a more general approximation,

$$\widehat{Z} = \frac{1}{N}\sum_{n=1}^{N} \frac{\pi(\mathbf{x}_n)}{\widehat{\pi}(\mathbf{x}_n)}. \tag{18}$$

Remark. It is very important to note that the estimator above is *biased* depending on the choices of **(a)** of the points $\mathbf{x}_1, ..., \mathbf{x}_N$, **(b)** the scale parameter h, and **(c)** the number of samples M for building $\widehat{\pi}(\mathbf{x})$. A improved version of this approximation can be obtained by the *importance sampling* approach described below, where $\mathbf{x}_1, ..., \mathbf{x}_N$ are drawn from the KDE mixture $\widehat{\pi}(\mathbf{x})$. In this case, the resulting estimator is *unbiased.*

Chib's method. In [11, 12], the authors present more sophisticated methods to estimate $\bar{\pi}(\mathbf{x}^*)$ using outputs from Gibbs sampling and the Metropolis-Hastings (MH) algorithm respectively [74]. Here we only present the latter method, since it can be applied in more general settings.
In [12], the authors propose to estimate the value of the posterior in one point $\mathbf{x}^*$, i.e., $\bar{\pi}(\mathbf{x}^*)$, using the output from a MH sampler. More specifically, let us denote the current state as $\mathbf{x}$. To draw a possible candidate as future state $\mathbf{z} \sim \varphi(\mathbf{z}|\mathbf{x})$ (where $\varphi(\mathbf{z}|\mathbf{x})$ the proposal density used within MH), then the probability of accepting the new state in a MH scheme is [51, 74]

$$\alpha(\mathbf{x}, \mathbf{z}) = \min\left\{1, \frac{\pi(\mathbf{z})\varphi(\mathbf{x}|\mathbf{z})}{\pi(\mathbf{x})\varphi(\mathbf{z}|\mathbf{x})}\right\}. \tag{19}$$

Note that this α satisfies the detailed balance condition, i.e.,

$$\alpha(\mathbf{x}, \mathbf{z})\varphi(\mathbf{z}|\mathbf{x})\bar{\pi}(\mathbf{x}) = \alpha(\mathbf{z}, \mathbf{x})\varphi(\mathbf{x}|\mathbf{z})\bar{\pi}(\mathbf{z}). \tag{20}$$

By integrating in $\mathbf{x}$ both sides, we obtain

$$\int_{\mathcal{X}} \alpha(\mathbf{x}, \mathbf{z})\varphi(\mathbf{z}|\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x} = \int_{\mathcal{X}} \alpha(\mathbf{z}, \mathbf{x})\varphi(\mathbf{x}|\mathbf{z})\bar{\pi}(\mathbf{z})d\mathbf{x} \tag{21}$$

$$\int_{\mathcal{X}} \alpha(\mathbf{x}, \mathbf{z})\varphi(\mathbf{z}|\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x} = \bar{\pi}(\mathbf{z})\int_{\mathcal{X}} \alpha(\mathbf{z}, \mathbf{x})\varphi(\mathbf{x}|\mathbf{z})d\mathbf{x}, \tag{22}$$

hence finally we can solve with respect to $\bar{\pi}(\mathbf{z})$ obtaining

$$\bar{\pi}(\mathbf{z}) = \frac{\int_{\mathcal{X}} \alpha(\mathbf{x}, \mathbf{z})\varphi(\mathbf{z}|\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x}}{\int_{\mathcal{X}} \alpha(\mathbf{z}, \mathbf{x})\varphi(\mathbf{x}|\mathbf{z})d\mathbf{x}}. \tag{23}$$

This suggests the following estimate of $\bar{\pi}(\mathbf{x}^*)$ at a specific point $\mathbf{x}^*$ (note that $\mathbf{x}^*$ plays the role of $\mathbf{z}$ in the equation above),

$$\widehat{\pi}(\mathbf{x}^*) = \frac{\frac{1}{N_1}\sum_{i=i}^{N_1} \alpha(\mathbf{x}_i, \mathbf{x}^*)\varphi(\mathbf{x}^*|\mathbf{x}_i)}{\frac{1}{N_2}\sum_{j=1}^{N_2} \alpha(\mathbf{x}^*, \mathbf{v}_j)}, \qquad \{\mathbf{x}_i\}_{i=1}^{N_1} \sim \bar{\pi}(\mathbf{x}), \;\; \{\mathbf{v}_j\}_{j=1}^{N_2} \sim \varphi(\mathbf{x}|\mathbf{x}^*). \tag{24}$$

The same outputs of the MH scheme can be considered as $\{\mathbf{x}_i\}_{i=1}^{N_1}$. The final estimator is again $\widehat{Z} = \frac{\pi(\mathbf{x}^*)}{\widehat{\pi}(\mathbf{x}^*)}$, i.e.,

$$\widehat{Z} = \frac{\pi(\mathbf{x}^*)\frac{1}{N_2}\sum_{j=1}^{N_2} \alpha(\mathbf{x}^*, \mathbf{v}_j)}{\frac{1}{N_1}\sum_{i=i}^{N_1} \alpha(\mathbf{x}_i, \mathbf{x}^*)\varphi(\mathbf{x}^*|\mathbf{x}_i)}, \qquad \{\mathbf{x}_i\}_{i=1}^{N_1} \sim \bar{\pi}(\mathbf{x}), \;\; \{\mathbf{v}_j\}_{j=1}^{N_2} \sim \varphi(\mathbf{x}|\mathbf{x}^*). \tag{25}$$

The point $\mathbf{x}^*$ is usually chosen in an high probability region. Interesting discussions are contained in [64, 65], where the authors also show that this estimator is a special case of bridge sampling idea described in Section 3.2.

Interpolative approaches. Another possibility is to approximate Z by substituting the true $\pi(\mathbf{x})$ with an interpolative or a regression function $\widehat{\pi}(\mathbf{x})$ in the integral (4). For simplicity, we focus on the interpolation case, but all the considerations can be easily extended for a regression scenario. Given a set of nodes $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathcal{X}$ and N nonlinear functions $k(\mathbf{x}, \mathbf{x}') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ chosen in advance by the user (generally, centered around $\mathbf{x}'$), we can build the interpolant of unnormalized posterior $\pi(\mathbf{x})$ as follows

$$\widehat{\pi}_u(\mathbf{x}) = \sum_{i=1}^{N} \beta_i k(\mathbf{x}, \mathbf{x}_i), \tag{26}$$

where $\beta_i \in \mathbb{R}$ and the subindex u denotes that is an approximation of unnormalized function $\pi(\mathbf{x})$. The coefficients β_i are chosen such that $\widehat{\pi}(\mathbf{x})$ interpolates the points $\{\mathbf{x}_n, \pi(\mathbf{x}_n)\}$, that is, $\widehat{\pi}(\mathbf{x}_n) = \pi(\mathbf{x}_n)$. Then, we desire that

$$\sum_{i=1}^{N} \beta_i k(\mathbf{x}_n, \mathbf{x}_i) = \pi(\mathbf{x}_n),$$

for all $n = 1, ..., N$. Hence, we can write a $N \times N$ linear system where the β_i are the N unknowns, i.e.,

$$\begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_N) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & k(\mathbf{x}_N, \mathbf{x}_2) & \dots & k(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix} = \begin{pmatrix} \pi(\mathbf{x}_1) \\ \pi(\mathbf{x}_2) \\ \vdots \\ \pi(\mathbf{x}_N) \end{pmatrix} \tag{27}$$

In matrix form, we have

$$\mathbf{K}\boldsymbol{\beta} = \mathbf{y}, \tag{28}$$

where $(\mathbf{K})_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{y} = [\pi(\mathbf{x}_1), \ldots, \pi(\mathbf{x}_N)]^\top$. Thus, the solution is $\boldsymbol{\beta} = \mathbf{K}^{-1}\mathbf{y}$. Now the interpolant $\widehat{\pi}_u(\mathbf{x}) = \sum_{i=1}^N \beta_i k(\mathbf{x}, \mathbf{x}_i)$ can be used to approximate Z as follows

$$\widehat{Z} = \int_{\mathcal{X}} \widehat{\pi}_u(\mathbf{x}) d\mathbf{x} = \sum_{i=1}^N \beta_i \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}_i) d\mathbf{x}. \tag{29}$$

If we are able to compute analytically $\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}_i) d\mathbf{x}$, we have an approximation $\widehat{Z}$. Some suitable choices of $k(\cdot, \cdot)$ are rectangular, triangular and Gaussian functions. More specifically, if all the nonlinearities $k(\mathbf{x}, \mathbf{x}_i)$ are normalized, the approximation of Z is $\widehat{Z} = \sum_{i=1}^N \beta_i$. This approach is related to the so-called Bayesian quadrature (using Gaussian process approximation) [73] and the sticky proposal constructions within MCMC or rejection sampling algorithms [32, 30, 50, 62]. Adaptive schemes adding sequentially more nodes could be also considered, improving the approximation $\widehat{Z}$ [30, 50].

3 Techniques based on IS

Most of the techniques for approximating the marginal likelihood are based on the importance sampling (IS) approach. Other methods are directly or indirectly related to the IS framework. In this sense, this section is the core of this survey. The standard IS scheme relies on the following equality,

$$Z = \int_{\mathcal{X}} \pi(\mathbf{x}) d\mathbf{x} = \int_{\mathcal{X}} \frac{\pi(\mathbf{x})}{\bar{q}(\mathbf{x})} \bar{q}(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{\bar{q}}\left[\frac{\pi(\mathbf{x})}{\bar{q}(\mathbf{x})}\right] \tag{30}$$

$$= \int_{\mathcal{X}} \frac{\ell(\mathbf{y}|\mathbf{x}) g(\mathbf{x})}{\bar{q}(\mathbf{x})} \bar{q}(\mathbf{x}) d\mathbf{x}, \tag{31}$$

where $\bar{q}(\mathbf{x})$ is a simpler normalized proposal density, $\int_{\mathcal{X}} \bar{q}(\mathbf{x}) d\mathbf{x} = 1$. Drawing N independent samples from proposal $\bar{q}(\mathbf{x})$, the *unbiased* IS estimator of Z is

$$\widehat{Z}_{IS1} = \frac{1}{N} \sum_{i=1}^N \frac{\pi(\mathbf{x}_i)}{\bar{q}(\mathbf{x}_i)} \tag{32}$$

$$= \frac{1}{N} \sum_{i=1}^N w_i, \tag{33}$$

$$= \frac{1}{N} \sum_{i=1}^N \frac{\ell(\mathbf{y}|\mathbf{x}_i) g(\mathbf{x}_i)}{\bar{q}(\mathbf{x}_i)} = \frac{1}{N} \sum_{i=1}^N \rho_i \ell(\mathbf{y}|\mathbf{x}_i), \qquad \{\mathbf{x}_i\}_{i=1}^N \sim \bar{q}(\mathbf{x}), \tag{34}$$

where $w_i = \frac{\pi(\mathbf{x}_i)}{\bar{q}(\mathbf{x}_i)}$ are the standard IS weights and $\rho_i = \frac{g(\mathbf{x}_i)}{\bar{q}(\mathbf{x}_i)}$. An alternative IS estimator (denoted as IS vers-2) is given by, considering a possibly unnormalized proposal pdf $q(\mathbf{x}) \propto \bar{q}(\mathbf{x})$ (the case

$q(\mathbf{x}) = \bar{q}(\mathbf{x})$ is also included),

$$\widehat{Z}_{IS2} = \frac{1}{\sum_{n=1}^{N} \frac{g(\mathbf{x}_n)}{q(\mathbf{x}_n)}} \sum_{i=1}^{N} \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)} \ell(\mathbf{y}|\mathbf{x}_i), \tag{35}$$

$$= \frac{1}{\sum_{n=1}^{N} \rho_n} \sum_{i=1}^{N} \rho_i \ell(\mathbf{y}|\mathbf{x}_i), \tag{36}$$

$$= \sum_{i=1}^{N} \bar{\rho}_i \ell(\mathbf{y}|\mathbf{x}_i), \qquad \{\mathbf{x}_i\}_{i=1}^{N} \sim \bar{q}(\mathbf{x}). \tag{37}$$

The estimator above is biased. However, it is a convex combination of likelihood values $\ell(\mathbf{y}|\mathbf{x}_i)$ since $\sum_{i=1}^{N} \bar{\rho}_i = 1$. Hence, in this case $\min_i \ell(\mathbf{y}|\mathbf{x}_i) \leq \widehat{Z} \leq \max_i \ell(\mathbf{y}|\mathbf{x}_i)$. Moreover, the estimator allows the use of an unnormalized proposal pdf $q(\mathbf{x}) \propto \bar{q}(\mathbf{x})$ and $\rho_i = \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)}$. For instance, one could consider $\bar{q}(\mathbf{x}) = \bar{\pi}(\mathbf{x})$, i.e., generate samples $\{\mathbf{x}_i\}_{i=1}^{N} \sim \bar{\pi}(\mathbf{x})$ by an MCMC algorithm and then evaluate $\rho_i = \frac{g(\mathbf{x}_i)}{\pi(\mathbf{x}_i)}$. Table 2 summarizes the IS estimators and shows some important special cases that will be described in the next section.

Table 2: IS estimators Eqs. (32)-(35) and relevant special cases.

$$\widehat{Z}_{IS1} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)} \ell(\mathbf{y}|\mathbf{x}_i) = \frac{1}{N} \sum_{i=1}^{N} \rho_i \ell(\mathbf{y}|\mathbf{x}_i)$$

Name	Estimator	$q(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Need of MCMC	Unbiased
Naive Monte Carlo	$\frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{y}|\mathbf{x}_i)$	$g(\mathbf{x})$	$g(\mathbf{x})$	—	✓

$$\widehat{Z}_{IS2} = \frac{1}{\sum_{n=1}^{N} \frac{g(\mathbf{x}_n)}{q(\mathbf{x}_n)}} \sum_{i=1}^{N} \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)} \ell(\mathbf{y}|\mathbf{x}_i) = \sum_{i=1}^{N} \bar{\rho}_i \ell(\mathbf{y}|\mathbf{x}_i)$$

Name	Estimator	$q(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Need of MCMC	Unbiased
Naive Monte Carlo	$\frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{y}|\mathbf{x}_i)$	$g(\mathbf{x})$	$g(\mathbf{x})$	—	✓
Harmonic mean	$\left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)}\right)^{-1}$	$\pi(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	✓	—

Two different sub-families of IS schemes are commonly used for computing normalizing constants (see also [10, chapter 5]). The first approach uses draws from a proposal density $\bar{q}(\mathbf{x})$ that is completely known (i.e. direct sampling and evaluate). Sophisticated choices of $\bar{q}(\mathbf{x})$ frequently imply the use of MCMC algorithms to sample from $\bar{q}(\mathbf{x})$ and that we can only evaluate $q(\mathbf{x}) \propto \bar{q}(\mathbf{x})$. The second class is formed by methods which uses more than one proposal density. We describe the methods in an increasing order of complexity.

3.1 Techniques using draws from one proposal density

In this section, all the techniques are IS schemes which use a unique proposal pdf, and are based on the identity Eq. (30). The techniques differ for the choice of $\bar{q}(\mathbf{x})$. Note that the optimal proposal choice for IS should be $\bar{q}(\mathbf{x}) = \bar{\pi}(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x})$. This choice is clearly difficult for two reasons: (a) we have to draw from $\bar{\pi}$ and (b) we do not know Z, hence we cannot evaluate $\bar{q}(\mathbf{x})$ but only $q(\mathbf{x}) = \pi(\mathbf{x})$ (where $q(\mathbf{x}) \propto \bar{q}(\mathbf{x})$). However, there are some methods based on this idea, as shown below.

Naive Monte Carlo (arithmetic mean estimator). It is straightforward to note that the integral above can be expressed as $Z = \mathbb{E}_g[\ell(\mathbf{y}|\mathbf{x})]$, then we can draw N samples $\{\mathbf{x}_i\}_{i=1}^N$ from the prior $g(\mathbf{x})$) and compute the following estimator

$$\widehat{Z} = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{y}|\mathbf{x}_i), \qquad \{\mathbf{x}_i\}_{i=1}^N \sim g(\mathbf{x}). \tag{38}$$

Namely a simple average of the likelihoods of a sample from the prior. Note that $\widehat{Z}$ will be very inefficient (large variance) if the posterior is much more concentrated than the prior (i.e., small overlap between likelihood and prior pdfs). Therefore, alternatives have been proposed, see below. It is a special case of the IS estimator with the choice $\bar{q}(\mathbf{x}) = g(\mathbf{x})$ (i.e., the proposal pdf is the prior).

Self-normalized Importance Sampling (Self-IS). Let us consider that the proposal is not normalized, and we can evaluate it up to a normalizing constant $q(\mathbf{x}) \propto \bar{q}(\mathbf{x})$. We also denote $c = \int_{\mathcal{X}} q(\mathbf{x})d\mathbf{x}$. Note that this also occurs in the ideal case of using $\bar{q}(\mathbf{x}) = \bar{\pi}(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x})$ where $c = Z$ and $q(\mathbf{x}) = \pi(\mathbf{x})$. In this case, we have

$$\frac{\widehat{Z}}{c} = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi(\mathbf{x}_i)}{q(\mathbf{x}_i)}, \qquad \{\mathbf{x}_i\}_{i=1}^N \sim \bar{q}(\mathbf{x}). \tag{39}$$

Therefore, we need an additional estimation of c. We can also use IS for this goal, considering a new normalized reference function $f(\mathbf{x})$, i.e., $\int_{\mathcal{X}} f(\mathbf{x})d\mathbf{x} = 1$. Now,

$$\frac{1}{c} = E_{\bar{q}}\left[\frac{f(\mathbf{x})}{q(\mathbf{x})}\right] = \int_{\mathcal{X}} \frac{f(\mathbf{x})}{q(\mathbf{x})}\bar{q}(\mathbf{x})d\mathbf{x} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{x}_i)}{q(\mathbf{x}_i)}, \qquad \{\mathbf{x}_i\}_{i=1}^N \sim \bar{q}(\mathbf{x}). \tag{40}$$

Thus, the *self-normalized* IS estimator is

$$\widehat{Z} = \frac{1}{\sum_{i=1}^{N} \frac{f(\mathbf{x}_i)}{q(\mathbf{x}_i)}} \sum_{i=1}^{N} \frac{\pi(\mathbf{x}_i)}{q(\mathbf{x}_i)}. \qquad \{\mathbf{x}_i\}_{i=1}^N \sim \bar{q}(\mathbf{x}). \tag{41}$$

We show later that the self-normalized IS estimator is strictly related to the umbrella sampling idea described below, and contains as special cases the rest of estimators included in this section (see Table 5).

Reverse Importance Sampling (RIS). The RIS scheme [27], also known *reciprocal* IS, can be derived from the identity

$$\frac{1}{Z}=\mathbb{E}_{\bar{\pi}}\left[\frac{f(\mathbf{x})}{\pi(\mathbf{x})}\right] \quad = \quad \int_{\mathcal{X}}\frac{f(\mathbf{x})}{\pi(\mathbf{x})}\bar{\pi}(\mathbf{x})d\mathbf{x} \tag{42}$$

where we consider an auxiliary normalized function $f(\mathbf{x})$, i.e., $\int_{\mathcal{X}} f(\mathbf{x})d\mathbf{x}=1$. Then, one could consider the estimator

$$\widehat{Z} \quad = \quad \left(\frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{x}_i)}{\pi(\mathbf{x}_i)}\right)^{-1} \quad = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{x}_i)}{\ell(\mathbf{y}|\mathbf{x}_i)g(\mathbf{x}_i)}\right)^{-1}, \quad \mathbf{x}_i\sim\bar{\pi}(\mathbf{x}) \text{ (via MCMC).} \tag{43}$$

The estimator above is consistent but biased. Indeed, the expression $\frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{x}_i)}{\pi(\mathbf{x}_i)}$ is a unbiased estimator of $1/Z$, but $\widehat{Z}$ in the Eq. (43) is not an unbiased estimator of Z. Note that $\bar{\pi}(\mathbf{x})$ plays the role of importance density from which we need to draw from. Therefore, another sampling technique must be used (such as a MCMC method) in order to generated samples from $\bar{\pi}(\mathbf{x})$. In this case, we do not need samples from $f(\mathbf{x})$, although its choice affects the precision of the approximation. Unlike, in the standard IS approach, $f(\mathbf{x})$ must have lighter tails than $\pi(\mathbf{x})=\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})$. For further details, see the example in Section 7.1. The RIS estimator is a special case of self-normalized IS estimator in Eq. (41) when $\bar{q}(\mathbf{x})=\bar{\pi}(\mathbf{x})$ and $q(\mathbf{x})=\pi(\mathbf{x})$.

Harmonic mean (HM) estimators. The HM estimator can be directly derived from the following expected value,

$$\mathbb{E}_{\bar{\pi}}\left[\frac{1}{\ell(\mathbf{y}|\mathbf{x})}\right]=\int_{\mathcal{X}}\frac{1}{\ell(\mathbf{y}|\mathbf{x})}\bar{\pi}(\mathbf{x})d\mathbf{x}, \tag{44}$$

$$=\frac{1}{Z}\int_{\mathcal{X}}\frac{1}{\ell(\mathbf{y}|\mathbf{x})}\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})d\mathbf{x}=\frac{1}{Z}\int_{\mathcal{X}}g(\mathbf{x})d\mathbf{x}=\frac{1}{Z}. \tag{45}$$

The main idea is again to use the posterior itself as proposal. Since direct sampling from $\bar{\pi}(\mathbf{x})$ is generally impossible, this task requires the use of MCMC algorithms. Thus, the HM estimator is

$$\widehat{Z}=\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)}}, \quad \{\mathbf{x}_i\}_{i=1}^{N}\sim\bar{\pi}(\mathbf{x}) \text{ (via MCMC).} \tag{46}$$

The estimator above is a special case of RIS when $f(\mathbf{x})=g(\mathbf{x})$ in Eq. (43) (i.e., the prior pdf). Moreover, the HM estimator is also a special case of Self-IS in Eq. (41), setting again $f(\mathbf{x})=g(\mathbf{x})$ and $\bar{q}(\mathbf{x})=\bar{\pi}(\mathbf{x})$ (so that $q(\mathbf{x})=\pi(\mathbf{x})$). The HM estimator converges almost surely to the correct value, but the variance of $\widehat{Z}^{-1}$ is often infinite. . This manifests itself by the occasional occurrence of a value of $\mathbf{x}_i$ with small likelihood and hence large effect, so that the estimator can be somewhat unstable.[1] Generally, the HM estimator tends to overestimate the marginal likelihood. For further

[1] See the comments of Radford Neal's blog, `https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/`, where R. Neal defines the HM estimator as "the worst estimator ever".

details, see the example in Section 7.1 recalling that the HM estimator is a special case of RIS.

Summary 1. Table 3 summarizes the estimators described above. Note that in the standard IS estimator the option $\bar{q}(\mathbf{x}) = \bar{\pi}(\mathbf{x})$ is not feasible, whereas it is possible for its second version.

Table 3: One-proposal estimators of Z

Name	Estimator	Proposal pdf	Need of MCMC	Unbiased
Standard IS	$\frac{1}{N}\sum_{i=1}^N \rho_i \ell(\mathbf{y}\mid\mathbf{x}_i)$	Generic, $\bar{q}(\mathbf{x})$	—	✓
Standard IS vers-2	$\sum_{i=1}^N \bar{\rho}_i \ell(\mathbf{y}\mid\mathbf{x}_i)$	Generic, $\bar{q}(\mathbf{x})$	no, if $\bar{q}(\mathbf{x}) \neq \bar{\pi}(\mathbf{x})$	—
Naive Monte Carlo	$\frac{1}{N}\sum_{i=1}^N \ell(\mathbf{y}\mid\mathbf{x}_i)$	Prior, $g(\mathbf{x})$	—	✓
Harmonic mean	$\left(\frac{1}{N}\sum_{i=1}^N \frac{1}{\ell(\mathbf{y}\mid\mathbf{x}_i)}\right)^{-1}$	Posterior, $\bar{\pi}(\mathbf{x})$	✓	—
RIS	$\left(\frac{1}{N}\sum_{i=1}^N \frac{f(\mathbf{x}_i)}{\pi(\mathbf{x}_i)}\right)^{-1}$	Posterior, $\bar{\pi}(\mathbf{x})$	✓	—
Self-IS	$\left(\sum_{i=1}^N \frac{f(\mathbf{x}_i)}{q(\mathbf{x}_i)}\right)^{-1} \sum_{i=1}^N \frac{\pi(\mathbf{x}_i)}{q(\mathbf{x}_i)}$	Generic, $\bar{q}(\mathbf{x})$	no, if $\bar{q}(\mathbf{x}) \neq \bar{\pi}(\mathbf{x})$	—

Some of the estimators in Table 3 can be unified within a common formulation. Let us consider the problem of estimating ratios of two normalizing constants c_1/c_2, where $c_i = \int q_i(\mathbf{x})d\mathbf{x}$ and $\bar{q}_i(\mathbf{x}) = q_i(\mathbf{x})/c_i$, $i = 1, 2$. Assuming we can evaluate both $q_1(\mathbf{x}), q_2(\mathbf{x})$, and draw samples from one of them, say $\bar{q}_2(\mathbf{x})$, the importance sampling estimator of c_1/c_2 is

$$\frac{c_1}{c_2} = \mathbb{E}_{\bar{q}_2}\left[\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})}\right] \approx \frac{1}{N}\sum_{i=1}^N \frac{q_1(\mathbf{x}_i)}{q_2(\mathbf{x}_i)}, \quad \{\mathbf{x}_i\}_{i=1}^N \sim \bar{q}_2(\mathbf{x}). \tag{47}$$

This framework includes the estimators discussed in this section, which are based on simulating from just one proposal density, as shown in Table 4.

Table 4: Summary of techniques considering the expression (47).

Name	$q_1(\mathbf{x})$	$q_2(\mathbf{x})$	c_1	c_2	Proposal pdf $\bar{q}_2(\mathbf{x})$	Estimator of c_1/c_2
Standard IS	$\pi(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	$\bar{q}(\mathbf{x})$	Z
Naive Monte Carlo	$\pi(\mathbf{x})$	$g(\mathbf{x})$	Z	1	$g(\mathbf{x})$	Z
Harmonic mean	$g(\mathbf{x})$	$\pi(\mathbf{x})$	1	Z	$\bar{\pi}(\mathbf{x})$	$1/Z$
RIS	$f(\mathbf{x})$	$\pi(\mathbf{x})$	1	Z	$\bar{\pi}(\mathbf{x})$	$1/Z$

Below we consider an extension of Eq. (47) where an additional density $\bar{q}_3(\mathbf{x})$ is employed for generating samples.

Umbrella Sampling (a.k.a. ratio importance sampling). The IS estimator of c_1/c_2 given in Eq. (47) may be inefficient when there is little overlap between $\bar{q}_1(\mathbf{x})$ and $\bar{q}_2(\mathbf{x})$, i.e., when

$\int_{\mathcal{X}} \bar{q}_1(\mathbf{x})\bar{q}_2(\mathbf{x})d\mathbf{x}$ is small. Umbrella sampling (originally proposed in the computational physics literature, [86]; also studied under the name "ratio importance sampling" in [9]) is based on the identity

$$\frac{c_1}{c_2} = \frac{c_1/c_3}{c_2/c_3} = \frac{\mathbb{E}_{\bar{q}_3}\left[\frac{q_1(\mathbf{x})}{q_3(\mathbf{x})}\right]}{\mathbb{E}_{\bar{q}_3}\left[\frac{q_2(\mathbf{x})}{q_3(\mathbf{x})}\right]} \approx \frac{\sum_{i=1}^{N}\frac{q_1(\mathbf{x}_i)}{q_3(\mathbf{x}_i)}}{\sum_{i=1}^{N}\frac{q_2(\mathbf{x}_i)}{q_3(\mathbf{x}_i)}}, \quad \{\mathbf{x}_i\}_{i=1}^{N} \sim \bar{q}_3(\mathbf{x}) \tag{48}$$

where $\bar{q}_3(\mathbf{x}) \propto q_3(\mathbf{x})$ represents a "middle" density, which is constructed to have large overlaps with both $\bar{q}_i(\mathbf{x})$, $i = 1, 2$. The performance of umbrella sampling clearly depends on the choice of $\bar{q}_3(\mathbf{x})$. The optimal umbrella sampling density $\bar{q}_3^{\text{opt}}(\mathbf{x})$, that minimizes the asymptotic relative mean-square error, is

$$\bar{q}_3^{\text{opt}}(\mathbf{x}) = \frac{|\bar{q}_1(\mathbf{x}) - \bar{q}_2(\mathbf{x})|}{\int |\bar{q}_1(\mathbf{x}') - \bar{q}_2(\mathbf{x}')|d\mathbf{x}'} = \frac{|q_1(\mathbf{x}) - \frac{c_1}{c_2}q_2(\mathbf{x})|}{\int |q_1(\mathbf{x}') - \frac{c_1}{c_2}q_2(\mathbf{x}')|d\mathbf{x}'}. \tag{49}$$

Since this $\bar{q}_3^{\text{opt}}(\mathbf{x})$ depends on the unknown ratio $\frac{c_1}{c_2}$ is not available for a direct use. The following two-stage procedure is often used in practice:

1. *Stage 1*: Draw N_1 samples from an arbitrary density $\bar{q}_3^{(1)}(\mathbf{x})$ and use them to obtain

$$\widehat{r}^{(1)} = \frac{\sum_{i=1}^{N_1}\frac{q_1(\mathbf{x}_i)}{q_3^{(1)}(\mathbf{x}_i)}}{\sum_{i=1}^{N_1}\frac{q_2(\mathbf{x}_i)}{q_3^{(1)}(\mathbf{x}_i)}}, \quad \{\mathbf{x}_i\}_{i=1}^{N_1} \sim \bar{q}_3^{(1)}(\mathbf{x}). \tag{50}$$

   and define

$$\bar{q}_3^{(2)}(\mathbf{x}) \propto |q_1(\mathbf{x}) - \widehat{r}^{(1)}q_2(\mathbf{x})|. \tag{51}$$

2. *Stage 2*: Draw N_2 samples from $\bar{q}_3^{(2)}(\mathbf{x})$ via MCMC and define the umbrella sampling estimator $\widehat{r}^{(2)}$ of $\frac{c_1}{c_2}$ as follows

$$\widehat{r}^{(2)} = \frac{\sum_{i=1}^{n_2}\frac{q_1(\mathbf{x}_i)}{q_3^{(2)}(\mathbf{x}_i)}}{\sum_{i=1}^{n_2}\frac{q_2(\mathbf{x}_i)}{q_3^{(2)}(\mathbf{x}_i)}}, \quad \{\mathbf{x}_i\}_{i=1}^{n_2} \sim \bar{q}_3^{(2)}(\mathbf{x}). \tag{52}$$

Summary 2. Considering $q_1(\mathbf{x}) = \pi(\mathbf{x})$, $q_2(\mathbf{x}) = \bar{q}_2(x) = f(\mathbf{x})$, $c_1 = Z$, $c_2 = 1$ and $c_3 \in \mathbb{R}$ in Eq. (48), we obtain

$$\widehat{Z} = \frac{1}{\sum_{i=1}^{N}\frac{f(\mathbf{x}_i)}{q_3(\mathbf{x}_i)}}\sum_{i=1}^{N}\frac{\pi(\mathbf{x}_i)}{q_3(\mathbf{x}_i)}. \qquad \{\mathbf{x}_i\}_{i=1}^{N} \sim \bar{q}_3(\mathbf{x}). \tag{53}$$

which is the self-normalized IS (Self-IS) estimator in Eq. (53). Table 5 summarizes all the techniques obtained from the identity (48) for $\frac{c_1}{c_2} = Z$. Note that Self-IS has the more general form and includes the rest of estimators as special cases. In the next section, we discuss a generalization of Eq. (47) for the case where we use samples from both $\bar{q}_1(\mathbf{x})$ and $\bar{q}_2(\mathbf{x})$.

Table 5: Summary of techniques considering the umbrella sampling identity (48) for computing $\frac{c_1}{c_2} = Z$. Note that Self-IS has the more general form and includes the rest of estimators as special cases.

Name	$q_1(\mathbf{x})$	$q_2(\mathbf{x})$	$q_3(\mathbf{x})$	c_1	c_2	c_3	sampling from $\bar{q}_3(\mathbf{x})$
Self-IS	$\pi(\mathbf{x})$	$f(\mathbf{x})$	$q(\mathbf{x})$	Z	1	c_3	$\bar{q}(\mathbf{x})$
Naive Monte Carlo	$\pi(\mathbf{x})$	$g(\mathbf{x})$	$g(\mathbf{x})$	Z	1	1	$g(\mathbf{x})$
Harmonic Mean	$\pi(\mathbf{x})$	$g(\mathbf{x})$	$\pi(\mathbf{x})$	Z	1	Z	$\bar{\pi}(\mathbf{x})$
RIS	$\pi(\mathbf{x})$	$f(\mathbf{x})$	$\pi(\mathbf{x})$	Z	1	Z	$\bar{\pi}(\mathbf{x})$
Standard IS; Eq. (32)	$\pi(\mathbf{x})$	$\bar{q}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	1	$\bar{q}(\mathbf{x})$
Standard IS vers-2; Eq. (35)	$\pi(\mathbf{x})$	$g(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	1	$\bar{q}(\mathbf{x})$

3.2 Techniques using draws from two proposal densities

In the previous section we considered estimators of Z that use samples drawn from a single proposal density. In this section we consider estimators of Z that generate samples from two proposal densities, denoted as $\bar{q}_i(\mathbf{x}) = \dfrac{q_i(\mathbf{x})}{c_i}, i = 1, 2$. All the techniques, that we will describe below, are based on the following *bridge sampling* identity [63],

$$\frac{c_1}{c_2} = \frac{\mathbb{E}_{\bar{q}_2}[q_1(\mathbf{x})\alpha(\mathbf{x})]}{\mathbb{E}_{\bar{q}_1}[q_2(\mathbf{x})\alpha(\mathbf{x})]}. \tag{54}$$

Note that the expression above is an extension of the Eq. (47). Indeed, taking $\alpha(\mathbf{x}) = \frac{1}{q_2(\mathbf{x})}$, we recover Eq. (47). Moreover, If we set $q_1(\mathbf{x}) = \pi(\mathbf{x})$, $c_1 = Z$, $q_2(\mathbf{x}) = \bar{q}(\mathbf{x})$ and $c_2 = 1$, then the identity becomes

$$Z = \frac{\mathbb{E}_{\bar{q}}\left[\pi(\mathbf{x})\alpha(\mathbf{x})\right]}{\mathbb{E}_{\bar{\pi}}\left[\bar{q}(\mathbf{x})\alpha(\mathbf{x})\right]}. \tag{55}$$

Figure 1 summarizes the connections among the Eqs. (47), (48), (54), (55) and the corresponding different methods. The standard IS and RIS schemes have been described in the previous sections, whereas the corresponding *locally-restricted* versions will be introduced below.

Locally-restricted IS and RIS. In the literature, there exist variants of the estimators in Eqs. (38) and (46). These corrected estimators are attempts to improve the efficiency (e.g., remove the infinite variance cases, specially in the harmonic estimator) by restricting the integration to a smaller subset of $\mathcal{X}$ (usually chosen in high posterior/likelihood-valued regions) generally denoted by $\mathcal{B} \subset \mathcal{X}$. As an example, $\mathcal{B}$ can be rectangular or ellipsoidal region centered at the MAP estimate $\widehat{\mathbf{x}}_{\text{MAP}}$.

Locally-restricted IS estimator. Consider the posterior mass of subset $\mathcal{B}$

$$\bar{\Pi}(\mathcal{B}) = \int_{\mathcal{B}} \bar{\pi}(\mathbf{x})d\mathbf{x} = \int_{\mathcal{X}} \mathbb{I}_{\mathcal{B}}(\mathbf{x})\frac{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{Z}d\mathbf{x}, \tag{56}$$

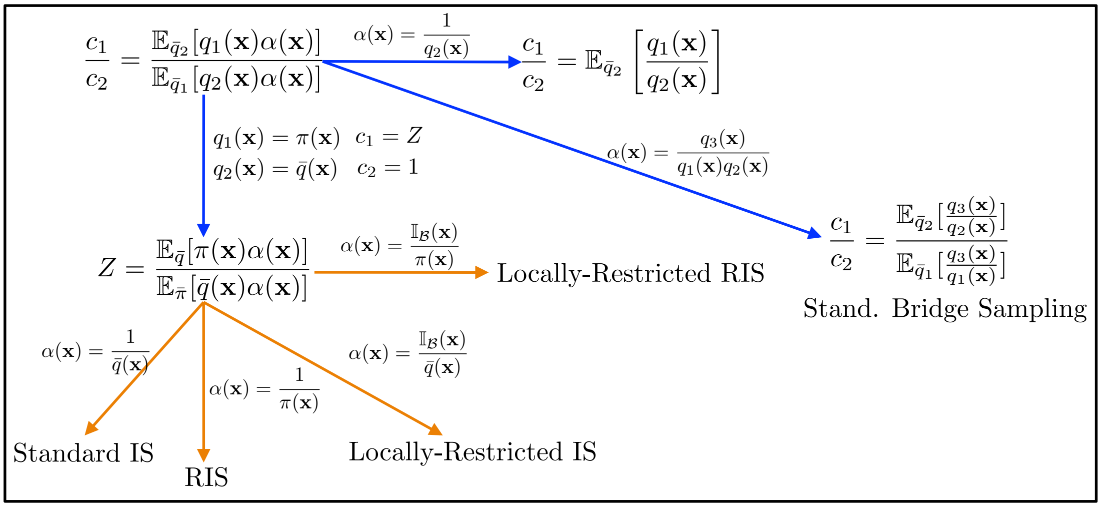

$\frac{c_1}{c_2} = \frac{\mathbb{E}_{\bar{q}_2}[q_1(\mathbf{x})\alpha(\mathbf{x})]}{\mathbb{E}_{\bar{q}_1}[q_2(\mathbf{x})\alpha(\mathbf{x})]}$
$\alpha(\mathbf{x}) = \frac{1}{q_2(\mathbf{x})}$
$\frac{c_1}{c_2} = \mathbb{E}_{\bar{q}_2}\left[\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})}\right]$
$q_1(\mathbf{x}) = \pi(\mathbf{x})$ $c_1 = Z$
$q_2(\mathbf{x}) = \bar{q}(\mathbf{x})$ $c_2 = 1$
$\alpha(\mathbf{x}) = \frac{q_3(\mathbf{x})}{q_1(\mathbf{x})q_2(\mathbf{x})}$
$\frac{c_1}{c_2} = \frac{\mathbb{E}_{\bar{q}_2}[\frac{q_3(\mathbf{x})}{q_2(\mathbf{x})}]}{\mathbb{E}_{\bar{q}_1}[\frac{q_3(\mathbf{x})}{q_1(\mathbf{x})}]}$
Stand. Bridge Sampling
$Z = \frac{\mathbb{E}_{\bar{q}}[\pi(\mathbf{x})\alpha(\mathbf{x})]}{\mathbb{E}_{\bar{\pi}}[\bar{q}(\mathbf{x})\alpha(\mathbf{x})]}$
$\alpha(\mathbf{x}) = \frac{\mathbb{I}_{\mathcal{B}}(\mathbf{x})}{\pi(\mathbf{x})}$
Locally-Restricted RIS
$\alpha(\mathbf{x}) = \frac{1}{\bar{q}(\mathbf{x})}$
$\alpha(\mathbf{x}) = \frac{1}{\pi(\mathbf{x})}$
$\alpha(\mathbf{x}) = \frac{\mathbb{I}_{\mathcal{B}}(\mathbf{x})}{\bar{q}(\mathbf{x})}$
Standard IS
RIS
Locally-Restricted IS

Figure 1: Graphical representation of the relationships among the Eqs. (47), (48), (54), (55) and the corresponding different methods.

where $\mathbb{I}_{\mathcal{B}}(\mathbf{x})$ is an indicator function, taking value 1 for $\mathbf{x} \in \mathcal{B}$ and 0 otherwise. It leads to the following representation

$$Z = \frac{1}{\bar{\Pi}(\mathcal{B})} \int_{\mathcal{X}} \mathbb{I}_{\mathcal{B}}(\mathbf{x})\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})d\mathbf{x} = \frac{1}{\bar{\Pi}(\mathcal{B})} \mathbb{E}_{\bar{q}}\left[\mathbb{I}_{\mathcal{B}}(\mathbf{x})\frac{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{\bar{q}(\mathbf{x})}\right]. \tag{57}$$

We can estimate $\bar{\Pi}(\mathcal{B})$ considering N_1 samples from $\bar{\pi}(\mathbf{x})$, and by taking the proportion of samples inside $\mathcal{B}$. The resulting locally-restricted IS estimator of Z

$$\widehat{Z} = \frac{\frac{1}{N_1}\sum_{i=1}^{N_1} \frac{\mathbb{I}_{\mathcal{B}}(\mathbf{z}_i)\ell(\mathbf{y}|\mathbf{z}_i)g(\mathbf{z}_i)}{\bar{q}(\mathbf{z}_i)}}{\frac{1}{N_2}\sum_{i=1}^{N_2} \mathbb{I}_{\mathcal{B}}(\mathbf{x}_i)}, \quad \{\mathbf{z}_i\}_{i=1}^{N_1} \sim \bar{q}(\mathbf{x}), \ \ \{\mathbf{x}_i\}_{i=1}^{N_2} \sim \bar{\pi}(\mathbf{x}) \quad \text{(via MCMC)}. \tag{58}$$

Note that the above estimator requires samples from two densities, namely the proposal $\bar{q}(\mathbf{x})$ and the posterior density $\bar{\pi}(\mathbf{x})$ (via MCMC).

Locally-restricted RIS estimator. To derive the locally-restricted RIS estimator, consider the mass of $\mathcal{B}$ under $\bar{q}(\mathbf{x})$

$$\bar{Q}(\mathcal{B}) = \int_{\mathcal{B}} \bar{q}(\mathbf{x})d\mathbf{x} = Z \cdot \mathbb{E}_{\bar{\pi}}\left[\mathbb{I}_{\mathcal{B}}(\mathbf{x})\frac{\bar{q}(\mathbf{x})}{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}\right], \tag{59}$$

which leads to the following representation

$$Z = \frac{\bar{Q}(\mathcal{B})}{\mathbb{E}_{\bar{\pi}}\left[\frac{\mathbb{I}_{\mathcal{B}}(\mathbf{x})\bar{q}(\mathbf{x})}{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}\right]}. \tag{60}$$

$\bar{Q}(\mathcal{B})$ can be estimated using a sample from $\bar{q}(\mathbf{x})$ by taking the proportion of sampled values inside $\mathcal{B}$. The locally-restricted RIS estimator is

$$\widehat{Z} = \frac{\frac{1}{N_1}\sum_{i=1}^{N_1} \mathbb{I}_{\mathcal{B}}(\mathbf{z}_i)}{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{\mathbb{I}_{\mathcal{B}}(\mathbf{x}_i)\bar{q}(\mathbf{x}_i)}{\ell(\mathbf{y}|\mathbf{x}_i)g(\mathbf{x}_i)}}, \quad \{\mathbf{z}_i\}_{i=1}^{N_1} \sim \bar{q}(\mathbf{x}), \qquad \{\mathbf{x}_i\}_{i=1}^{N_2} \sim \bar{\pi}(\mathbf{x}). \tag{61}$$

Other variants, where $\mathbb{I}_{\mathcal{B}}(\mathbf{z})$ corresponds to high posterior density regions, can be find in [75].

Optimal construction of bridge sampling. Identities as (54) are associated to the bridge sampling approach. However, considering $\alpha(\mathbf{x}) = \frac{\gamma(\mathbf{x})}{q_2(\mathbf{x})q_1(\mathbf{x})}$ in Eq. (54), bridge sampling can be also motivated from the expression

$$\frac{c_1}{c_2} = \frac{c_3/c_2}{c_3/c_1} = \frac{\mathbb{E}_{\bar{q}_2}\left[\frac{\gamma(\mathbf{x})}{q_2(\mathbf{x})}\right]}{\mathbb{E}_{\bar{q}_1}\left[\frac{\gamma(\mathbf{x})}{q_1(\mathbf{x})}\right]}, \tag{62}$$

where the density $\bar{\gamma}(\mathbf{x}) \propto \gamma(\mathbf{x})$ is in some sense "in between" $q_1(\mathbf{x})$ and $q_2(\mathbf{x})$. That is, instead of applying directly (47) to $\frac{c_1}{c_2}$, we apply it to first estimate $\frac{c_3}{c_2}$ and $\frac{c_3}{c_1}$ and then take the ratio to cancel c_3. The bridge sampling estimator of $\frac{c_1}{c_2}$ is then

$$\frac{c_1}{c_2} \approx \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{\gamma(\mathbf{z}_i)}{q_2(\mathbf{z}_i)}}{\frac{1}{N_1}\sum_{i=1}^{N_1} \frac{\gamma(\mathbf{x}_i)}{q_1(\mathbf{x}_i)}}, \quad \{\mathbf{x}_i\}_{i=1}^{N_1} \sim \bar{q}_1(\mathbf{x}), \quad \{\mathbf{z}_i\}_{i=1}^{N_2} \sim \bar{q}_2(\mathbf{x}). \tag{63}$$

We do not need to draw samples from $\bar{\gamma}(\mathbf{x})$, but only evaluate $\gamma(\mathbf{x})$. It can be shown that the optimal[2] bridge density $\bar{\gamma}(\mathbf{x})$ can be expressed as a weighted harmonic mean of $\bar{q}_1(\mathbf{x})$ and $\bar{q}_2(\mathbf{x})$ (with weights being the sampling rates),

$$\begin{aligned}
\bar{\gamma}^{\text{opt}}(\mathbf{x}) &= \frac{1}{\frac{N_2}{N_1+N_2}[\bar{q}_1(\mathbf{x})]^{-1} + \frac{N_1}{N_1+N_2}[\bar{q}_2(\mathbf{x})]^{-1}} && (64)\\
&= \frac{1}{c_2} \cdot \frac{1}{N_2\frac{c_1}{c_2}q_1^{-1}(\mathbf{x}) + N_1 q_2^{-1}(\mathbf{x})} && (65)\\
&\propto \gamma^{\text{opt}}(\mathbf{x}) = \frac{q_1(\mathbf{x})q_2(\mathbf{x})}{N_1 q_1(\mathbf{x}) + N_2\frac{c_1}{c_2}q_2(\mathbf{x})}, && (66)
\end{aligned}$$

depending on the unknown ratio $r = \frac{c_1}{c_2}$. Therefore, we cannot even evaluate $\gamma^{\text{opt}}(\mathbf{x})$. Hence, we need to resort to the following iterative procedure to approximate the optimal bridge sampling estimator. Noting that

$$\frac{\gamma^{\text{opt}}(\mathbf{x})}{q_2(\mathbf{x})} = \frac{q_1(\mathbf{x})}{N_1 q_1(\mathbf{x}) + rN_2 q_2(\mathbf{x})}, \qquad \frac{\gamma^{\text{opt}}(\mathbf{x})}{q_1(\mathbf{x})} = \frac{q_2(\mathbf{x})}{N_1 q_1(\mathbf{x}) + rN_2 q_2(\mathbf{x})}. \tag{67}$$

The iterative procedure is formed by the following steps:

[2] In the sense of providing the most efficient estimator of the ratio $\frac{c_1}{c_2}$.

1. Start with an initial estimate $\widehat{r}^{(1)} \approx \frac{c_1}{c_2}$ (using e.g. Laplace's).
2. For $t = 1, ..., T$:
    (a) Draw $\{\mathbf{x}_i\}_{i=1}^{N_1} \sim \bar{q}_1(\mathbf{x})$ and $\{\mathbf{z}_i\}_{i=1}^{N_2} \sim \bar{q}_2(\mathbf{x})$ and iterate

$$\widehat{r}^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{q_1(\mathbf{z}_i)}{N_1 q_1(\mathbf{z}_i) + N_2 \widehat{r}^{(t)} q_2(\mathbf{z}_i)}}{\frac{1}{N_1}\sum_{i=1}^{N_1} \frac{q_2(\mathbf{x}_i)}{N_1 q_1(\mathbf{x}_i) + N_2 \widehat{r}^{(t)} q_2(\mathbf{x}_i)}}. \tag{68}$$

Optimal bridge sampling for Z. Given the considerations above, an iterative bridge sampling estimator of Z is obtained by setting $q_1(\mathbf{x}) = \pi(\mathbf{x})$, $c_1 = Z$, $\bar{q}_2(\mathbf{x}) = \bar{q}(\mathbf{x})$, so that

$$\widehat{Z}^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{\pi(\mathbf{z}_i)}{N_1 \pi(\mathbf{z}_i) + N_2 Z^{(t)} \bar{q}(\mathbf{z}_i)}}{\frac{1}{N_1}\sum_{i=1}^{N_1} \frac{\bar{q}(\mathbf{x}_i)}{N_1 \pi(\mathbf{x}_i) + N_2 Z^{(t)} \bar{q}(\mathbf{x}_i)}}, \quad \{\mathbf{z}_i\}_{i=1}^{N_2} \sim \bar{q}(\mathbf{x}) \text{ and } \{\mathbf{x}_i\}_{i=1}^{N_1} \sim \bar{\pi}(\mathbf{x}). \tag{69}$$

for $t = 1, ..., T$. When $N_1 = 0$, that is when all samples are drawn from $\bar{q}(\mathbf{x})$, the estimator above reduces to (non-iterative) standard IS scheme with proposal $\bar{q}(\mathbf{x})$. When $N_2 = 0$, that is when all samples are drawn from $\pi(\mathbf{x})$, the estimator becomes the (non-iterative) RIS estimator. See [9] for a comparison of optimal umbrella sampling, bridge sampling and path sampling (described in the next section). An alternative derivation of the optimal bridge sampling estimator is given in [75], by simulating from a mixture of type $\psi(\mathbf{x}) \propto \pi(\mathbf{x}) + r\bar{q}(\mathbf{x})$. However, the resulting estimator employs the same samples drawn from $\psi(\mathbf{x})$ in the numerator and denominator, unlike in Eq. (69).

Several techniques described in the last two subsections, including both umbrella and bridge sampling, are encompassed by the generic formula

$$\frac{c_1}{c_2} = \mathbb{E}_{\bar{\xi}}[q_1(\mathbf{x})\alpha(\mathbf{x})] \Big/ \mathbb{E}_{\bar{\chi}}[q_2(\mathbf{x})\alpha(\mathbf{x})] \tag{70}$$

as shown in Table 6.

3.3 IS based on multiple proposal densities

In this section we consider estimators of Z using samples drawn from more than two proposal densities. Typically these densities form a sequence of functions that are in some sense "in the middle" between the posterior $\bar{\pi}(\mathbf{x})$ and an easier-to-work-with density (e.g. the prior $g(\mathbf{x})$ or some other proposal density). The number of such middle densities must be specified by the user, and in some cases, it is equivalent to the selection of a *temperature schedule* for linking $g(\mathbf{x})$ and $\bar{\pi}(\mathbf{x})$. The resulting pdfs are usually called as *tempered* posteriors and correspond to flatter, spread distributions. The use of the tempered pdfs usually improve the mixing of the

Table 6: Summary of the IS schemes (with one or two proposal pdfs), using Eq. (70).

$$\frac{c_1}{c_2} = \mathbb{E}_{\bar{\xi}}[q_1(\mathbf{x})\alpha(\mathbf{x})] \Big/ \mathbb{E}_{\bar{\chi}}[q_2(\mathbf{x})\alpha(\mathbf{x})]$$

Name	$\alpha(\mathbf{x})$	$\bar{\xi}(\mathbf{x})$	$\bar{\chi}(\mathbf{x})$	$q_1(\mathbf{x})$	$q_2(\mathbf{x})$	c_1	c_2	sampling from
Bridge Identity - Eq. (54)	$\alpha(\mathbf{x})$	$\bar{q}_2(\mathbf{x})$	$\bar{q}_1(\mathbf{x})$	$q_1(\mathbf{x})$	$q_2(\mathbf{x})$	c_1	c_2	$\bar{q}_1(\mathbf{x})$, $\bar{q}_2(\mathbf{x})$
Bridge Identity - Eq. (62)	$\frac{\gamma(\mathbf{x})}{q_2(\mathbf{x})q_1(\mathbf{x})}$	$\bar{q}_2(\mathbf{x})$	$\bar{q}_1(\mathbf{x})$	$q_1(\mathbf{x})$	$q_2(\mathbf{x})$	c_1	c_2	$\bar{q}_1(\mathbf{x})$, $\bar{q}_2(\mathbf{x})$
Identity - Eq. (47)	$\frac{1}{q_2(\mathbf{x})}$	$\bar{q}_2(\mathbf{x})$	$\bar{q}_1(\mathbf{x})$	$q_1(\mathbf{x})$	$q_2(\mathbf{x})$	c_1	c_2	$\bar{q}_2(\mathbf{x})$
Umbrella - Eq. (48)	$\frac{1}{q_3(\mathbf{x})}$	$\bar{q}_3(\mathbf{x})$	$\bar{q}_3(\mathbf{x})$	$q_1(\mathbf{x})$	$q_2(\mathbf{x})$	c_1	c_2	$\bar{q}_3(\mathbf{x})$
Self-norm. IS - Eqs. (41) (53)	$\frac{1}{q_3(\mathbf{x})}$	$\bar{q}_3(\mathbf{x})$	$\bar{q}_3(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$f(\mathbf{x})$	Z	1	$\bar{q}_3(\mathbf{x})$
Bridge Identity - Eq. (55)	$\alpha(\mathbf{x})$	$\bar{q}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	$\bar{\pi}(\mathbf{x})$, $\bar{q}(\mathbf{x})$
Standard IS	$1/\bar{q}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	$\bar{q}(\mathbf{x})$
RIS	$1/\bar{\pi}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	$\bar{\pi}(\mathbf{x})$
Locally-Restricted IS	$\mathbb{I}_{\mathcal{B}}(\mathbf{x})/\bar{q}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	$\bar{\pi}(\mathbf{x})$, $\bar{q}(\mathbf{x})$
Locally-Restricted RIS	$\mathbb{I}_{\mathcal{B}}(\mathbf{x})/\bar{\pi}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{\pi}(\mathbf{x})$	$\bar{q}(\mathbf{x})$	Z	1	$\bar{\pi}(\mathbf{x})$, $\bar{q}(\mathbf{x})$

algorithm and foster the exploration of the space $\mathcal{X}$. This idea is shared by path sampling, the power posterior methods and stepping-stone sampling described below. However, we start with a general IS scheme considering different proposals $\bar{q}_n(\mathbf{x})$'s. Some of them could be tempered posteriors and the generation would be performed by an MCMC method in this case.

Multiple Importance Sampling (MIS) estimators. Here, we consider to generate samples from different proposal densities, i.e.,

$$\mathbf{x}_n \sim \bar{q}_n(\mathbf{x}), \qquad n = 1, ..., N. \tag{71}$$

In this scenario, different proper importance weights can be used [24, 23, 22]. The most efficient MIS scheme considers the following weights

$$w_n = \frac{\pi(\mathbf{x}_n)}{\frac{1}{N}\sum_{i=1}^N \bar{q}_i(\mathbf{x}_n)} = \frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}, \tag{72}$$

where $\psi(\mathbf{x}_n) = \frac{1}{N}\sum_{i=1}^N \bar{q}_i(\mathbf{x}_n)$. Indeed, considering the set of samples $\{\mathbf{x}_n\}_{n=1}^N$ drawn in a deterministic order, $\mathbf{x}_n \sim \bar{q}_n(\mathbf{x})$, and given a sample $\mathbf{x}^* \in \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ uniformly chosen in $\{\mathbf{x}_n\}_{n=1}^N$, then we can write $\mathbf{x}^* \sim \psi(\mathbf{x}_n)$. The standard MIS estimator is

$$\begin{aligned}
\widehat{Z} = \frac{1}{N}\sum_{n=1}^N w_n \quad &= \quad \frac{1}{N}\sum_{n=1}^N \frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)} && (73) \\
&= \quad \frac{1}{N}\sum_{n=1}^N \frac{g(\mathbf{x}_n)\ell(\mathbf{y}|\mathbf{x}_n)}{\psi(\mathbf{x}_n)}, && (74) \\
&= \quad \frac{1}{N}\sum_{n=1}^N \eta_n \ell(\mathbf{y}|\mathbf{x}_n), \qquad \mathbf{x}_n \sim \bar{q}_n(\mathbf{x}), \qquad n = 1, ..., N. && (75)
\end{aligned}$$

where $\eta_n = \frac{g(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}$. The estimator is unbiased [24]. As in the standard IS scheme, an alternative biased estimator is

$$\widehat{Z} = \frac{1}{N}\sum_{n=1}^{N} \bar{\eta}_n \ell(\mathbf{y}|\mathbf{x}_n), \qquad \mathbf{x}_n \sim \bar{q}_n(\mathbf{x}), \qquad n = 1, ..., N, \tag{76}$$

where $\bar{\eta}_n = \frac{\eta_n}{\sum_{i=1}^N \eta_i}$, so that $\sum_{i=1}^N \bar{\eta}_i = 1$ and we have a convex combination of likelihood values $\ell(\mathbf{y}|\mathbf{x}_n)$'s.

Path sampling. The method of path sampling for estimating $\frac{c_1}{c_2}$ relies on the idea of building and drawing samples from a sequence of distributions linking $\bar{q}_1(\mathbf{x})$ and $\bar{q}_2(\mathbf{x})$ (a continuous path). For the purpose of estimating the marginal likelihood, we set $\bar{q}_2(\mathbf{x}) = g(\mathbf{x})$ and $\bar{q}_1(\mathbf{x}) = \bar{\pi}(\mathbf{x})$ and we "link them" by an univariate "path" with parameter β. Let

$$\pi(\mathbf{x}|\beta), \;\; \beta \in [0, 1], \tag{77}$$

denote a sequence of (probably unnormalized except for $\beta = 0$) densities such $\pi(\mathbf{x}|\beta = 0) = g(\mathbf{x})$ and $\pi(\mathbf{x}|\beta = 1) = \pi(\mathbf{x})$. The path sampling method for estimating the marginal likelihood is based on expressing $\log Z$ as

$$\log Z = \mathbb{E}_{p(\mathbf{x},\beta)}\left[\frac{U(\mathbf{x},\beta)}{p(\beta)}\right], \quad \text{with } U(\mathbf{x},\beta) = \frac{\partial}{\partial \beta}\log \pi(\mathbf{x}|\beta), \tag{78}$$

where the expectation is w.r.t. the joint $p(\mathbf{x},\beta) = \frac{\pi(\mathbf{x}|\beta)}{Z(\beta)}p(\beta)$, being $Z(\beta)$ the normalizing constant of $\pi(\mathbf{x}|\beta)$ and $p(\beta)$ represents a density for $\beta \in [0, 1]$. Indeed, we have

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x},\beta)}\left[\frac{U(\mathbf{x},\beta)}{p(\beta)}\right] &= \int_{\mathcal{X}}\int_0^1 \frac{1}{p(\beta)}\frac{\partial}{\partial\beta}\log\pi(\mathbf{x}|\beta)\frac{\pi(\mathbf{x}|\beta)}{Z(\beta)}p(\beta)d\mathbf{x}d\beta, && (79)\\
&= \int_{\mathcal{X}}\int_0^1 \frac{1}{\pi(\mathbf{x}|\beta)}\frac{\partial}{\partial\beta}\pi(\mathbf{x}|\beta)\frac{\pi(\mathbf{x}|\beta)}{Z(\beta)}d\mathbf{x}d\beta, && (80)\\
&= \int_{\mathcal{X}}\int_0^1 \frac{1}{Z(\beta)}\frac{\partial}{\partial\beta}\pi(\mathbf{x}|\beta)d\mathbf{x}d\beta, && (81)\\
&= \int_0^1 \frac{1}{Z(\beta)}\frac{\partial}{\partial\beta}\left(\int_{\mathcal{X}}\pi(\mathbf{x}|\beta)d\mathbf{x}\right)d\beta, && (82)\\
&= \int_0^1 \frac{1}{Z(\beta)}\frac{\partial}{\partial\beta}Z(\beta)d\beta, && (83)\\
&= \int_0^1 \frac{\partial}{\partial\beta}\log Z(\beta)d\beta, && (84)\\
&= \log Z(1) - \log Z(0) = \log Z, && (85)
\end{aligned}$$

where we substituted $Z(\beta = 1) = Z(1) = Z$ and $Z(\beta = 0) = Z(0) = 1$. Thus, using a sample $\{\mathbf{x}_i, \beta_i\}_{i=1}^N \sim p(\mathbf{x},\beta)$, we can write the path sampling estimator for $\log Z$

$$\widehat{\log Z} = \frac{1}{N}\sum_{i=1}^{N}\frac{U(\mathbf{x}_i,\beta_i)}{p(\beta_i)}, \quad \{\mathbf{x}_i,\beta_i\}_{i=1}^N \sim p(\mathbf{x},\beta). \tag{86}$$

The samples may be obtained by first drawing β' from $p(\beta)$ (see [28] for guidelines on optimal $p(\beta)$ in one dimension) and then applying some MCMC steps to draw from $\bar{\pi}(\mathbf{x}|\beta')$ given β'. Often the so-called geometric path is employed (see also [28] for extensions),

$$\pi(\mathbf{x}|\beta) = g(\mathbf{x})^{1-\beta}\pi(\mathbf{x})^{\beta} \tag{87}$$
$$= g(\mathbf{x})\ell(\mathbf{y}|\mathbf{x})^{\beta}, \quad \beta \in [0,1]. \tag{88}$$

Note that $\pi(\mathbf{x}|\beta)$ is the posterior with a powered, "less informative" -"wider" likelihood (for this reason, $\pi(\mathbf{x}|\beta)$ is often called a "power posterior"). In this case, we have

$$U(\mathbf{x},\beta) = \frac{\partial}{\partial\beta}\log\pi(\mathbf{x}|\beta) \tag{89}$$
$$= \log\ell(\mathbf{y}|\mathbf{x}), \tag{90}$$

so the path sampling identity becomes

$$\log Z = \mathbb{E}_{p(\mathbf{x},\beta)}\left[\frac{\log\ell(\mathbf{y}|\mathbf{x})}{p(\beta)}\right], \tag{91}$$

which is also used in the power posterior method of [25], described next.

Method of Power Posteriors. The previous expression (91) can also be converted into an integral in $[0,1]$ as follows

$$\log Z = \mathbb{E}_{p(\mathbf{x},\beta)}\left[\frac{\log\ell(\mathbf{y}|\mathbf{x})}{p(\beta)}\right], \tag{92}$$
$$= \int_0^1 d\beta \int_{\mathcal{X}} \frac{\log\ell(\mathbf{y}|\mathbf{x})}{p(\beta)}\frac{\pi(\mathbf{x}|\beta)}{Z(\beta)}p(\beta)d\mathbf{x}, \tag{93}$$
$$= \int_0^1 d\beta \int_{\mathcal{X}} \log\ell(\mathbf{y}|\beta)\frac{\pi(\mathbf{x}|\beta)}{Z(\beta)}d\mathbf{x}, \tag{94}$$
$$= \int_0^1 \mathbb{E}_{\bar{\pi}(\mathbf{x}|\beta)}\left[\log\ell(\mathbf{y}|\mathbf{x})\right]d\beta. \tag{95}$$

The power posterior method aims at estimating the integral above by applying a quadrature rule. For instance, using the trapezoidal rule, choosing a discretization $0 = \beta_0 < \beta_1 < \cdots < \beta_{I-1} < \beta_I = 1$, leads to an approximation of type

$$\widehat{\log Z} = \sum_{i=1}^{I-1}(\beta_{i+1}-\beta_i)\frac{\mathbb{E}_{\bar{\pi}(\mathbf{x}|\beta_{i+1})}\left[\log\ell(\mathbf{y}|\mathbf{x})\right] + \mathbb{E}_{\bar{\pi}(\mathbf{x}|\beta_i)}\left[\log\ell(\mathbf{y}|\mathbf{x})\right]}{2}, \tag{96}$$

where the expected values w.r.t. the power posteriors can be independently approximated via MCMC

$$\mathbb{E}_{\bar{\pi}(\mathbf{x}|\beta_i)}\left[\log\ell(\mathbf{y}|\mathbf{x})\right] \approx \frac{1}{N}\sum_{k=1}^{N}\log\ell(\mathbf{y}|\mathbf{x}_{k,i}), \quad \{\mathbf{x}_{k,i}\}_{k=1}^{N} \sim \bar{\pi}(\mathbf{x}|\beta_i), \quad i = 1,\ldots,I. \tag{97}$$

Power posteriors as proposal densities. Recall the general IS estimator of $Z = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})d\mathbf{x}$, which involves a weighted sum of likelihood evaluations at points $\{\mathbf{x}_i\}_{i=1}^N \sim \bar{q}(\mathbf{x})$ drawn from importance density $\bar{q}(\mathbf{x})$:

$$\widehat{Z} = \sum_{i=1}^{N} \bar{\rho}_i \ell(\mathbf{y}|\mathbf{x}_i), \quad \bar{\rho}_i = \frac{\frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)}}{\sum_{n=1}^{N} \frac{g(\mathbf{x}_n)}{q(\mathbf{x}_n)}} \propto \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)}, \tag{98}$$

where $\sum_{i=1}^N \bar{\rho}_i = 1$. Let $\bar{q}(\mathbf{x}) = \bar{\pi}(\mathbf{x}|\beta) \propto q(\mathbf{x}) = \pi(\mathbf{x}|\beta) = g(\mathbf{x})\ell(\mathbf{y}|\mathbf{x})^\beta$, that is we use the power posterior as importance density so $\bar{\rho}_i \propto \frac{g(\mathbf{x}_i)}{g(\mathbf{x}_i)\ell(\mathbf{y}|\mathbf{x}_i)^\beta} = \frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)^\beta}$. The resulting IS estimator is

$$\widehat{Z} = \frac{\sum_{i=1}^{N} \frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)^\beta} \ell(\mathbf{y}|\mathbf{x}_i)}{\sum_{i=1}^{N} \frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)^\beta}} \tag{99}$$

$$= \frac{\sum_{i=1}^{N} \ell(\mathbf{y}|\mathbf{x}_i)^{1-\beta}}{\sum_{i=1}^{N} \ell(\mathbf{y}|\mathbf{x}_i)^{-\beta}} \qquad \{\mathbf{x}_i\}_{i=1}^N \sim \bar{\pi}(\mathbf{x}|\beta) \quad \text{(via MCMC)}. \tag{100}$$

Table 7 shows that this technique includes different schemes for different values of β. Different possible MIS schemes can be also considered, i.e., using Eq. (76) for instance [24, 22].

Table 7: Different estimators of Z using $\bar{q}(\mathbf{x}) \propto g(\mathbf{x})\ell(\mathbf{y}|\mathbf{x})^\beta$ as importance density, with $\beta \in [0, 1]$.

Name	Coefficient β	Weights $\bar{\rho}_i$	Estimator $\widehat{Z} = \sum_{i=1}^N \bar{\rho}_i \ell(\mathbf{y}|\mathbf{x}_i)$
Naive Monte Carlo	$\beta = 0$	$\frac{1}{N}$	$\frac{1}{N}\sum_{i=1}^N \ell(\mathbf{y}|\mathbf{x}_i)$
Harmonic Mean Estimator	$\beta = 1$	$\frac{\frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)}}{\sum_j \frac{1}{\ell(\mathbf{y}|\mathbf{x}_j)}}$	$\widehat{Z} = \frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)}}$
Power posterior as proposal pdf	$0 < \beta < 1$	$\frac{\frac{1}{\ell(\mathbf{y}|\mathbf{x}_i)^\beta}}{\sum_j \frac{1}{\ell(\mathbf{y}|\mathbf{x}_j)^\beta}}$	$\widehat{Z} = \frac{\sum_i \ell(\mathbf{y}|\mathbf{x}_i)^{1-\beta}}{\sum_i \ell(\mathbf{y}|\mathbf{x}_i)^{-\beta}}$

Stepping-stone (SS) sampling. Consider again $\bar{\pi}(\mathbf{x}|\beta) \propto g(\mathbf{x})\ell(\mathbf{y}|\mathbf{x})^\beta$. The goal is to estimate $Z = \frac{Z(1)}{Z(0)}$, which can be expressed as the following product

$$Z = \frac{Z(1)}{Z(0)} = \prod_{k=1}^{K} \frac{Z(\beta_k)}{Z(\beta_{k-1})}, \tag{101}$$

where β_k are often chosen as $\beta_k = \frac{k}{K}$, $k = 1, \ldots, K$, i.e., with a uniform grid in $[0, 1]$. The idea of

SS sampling is to estimate each ratio $r_k = \frac{Z(\beta_k)}{Z(\beta_{k-1})}$ by importance sampling as

$$r_k = \frac{Z(\beta_k)}{Z(\beta_{k-1})} = \mathbb{E}_{\bar{\pi}(\mathbf{x}|\beta_{k-1})}\left[\frac{\pi(\mathbf{x}|\beta_k)}{\pi(\mathbf{x}|\beta_{k-1})}\right] \tag{102}$$

$$= \mathbb{E}_{\bar{\pi}(\mathbf{x}|\beta_{k-1})}\left[\frac{\ell(\mathbf{y}|\mathbf{x})^{\beta_k}}{\ell(\mathbf{y}|\mathbf{x})^{\beta_{k-1}}}\right] \tag{103}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{y}|\mathbf{x}_i)^{\beta_k-\beta_{k-1}}, \quad \{\mathbf{x}_i\}_{i=1}^{N} \sim \bar{\pi}(\mathbf{x}|\beta_{k-1}). \tag{104}$$

We can improve the numerical stability by factoring the largest sampled likelihood term $\ell_{\max} = \max_i \ell(\mathbf{y}|\mathbf{x}_i)$, so

$$\widehat{r}_k = \frac{1}{N}(\ell_{\max})^{\beta_k-\beta_{k-1}} \sum_{i=1}^{N} \left(\frac{\ell(\mathbf{y}|\mathbf{x}_i)}{\ell_{\max}}\right)^{\beta_k-\beta_{k-1}}, \quad \{\mathbf{x}_i\}_{i=1}^{N} \sim \bar{\pi}(\mathbf{x}|\beta_{k-1}). \tag{105}$$

Multiplying all ratio estimates yields the final estimator of Z

$$\widehat{Z} = \prod_{k=1}^{K} \widehat{r}_k. \tag{106}$$

Some strategies for selecting the values of β's are discussed in [25].

4 Advanced schemes combining MCMC and IS

In the previous sections, we have already introduced several methods which require the use of MCMC algorithms in order to draw from complex proposal densities. The RIS estimator, path sampling, power posteriors and the SS sampling schemes are some examples. In this section, we describe more sophisticated scheme which combines MCMC and IS techniques.

4.1 MCMC within IS schemes

In this section, we will see how to *properly* weight samples obtained by different MCMC iterations. We denote as $K(\mathbf{z}|\mathbf{x})$ the transition kernel which summarizes all the steps of the employed MCMC algorithm. Note that generally $K(\mathbf{z}|\mathbf{x})$ cannot be evaluated. However, we can use MCMC kernels $K(\mathbf{z}|\mathbf{x})$ in the same fashion as proposal densities, considering the concept of the so-called *proper weighting* [45, 53].

4.1.1 Weighting a sample after an MCMC iteration

Let us consider the following procedure:

1. Draw $\mathbf{x}_0 \sim q(\mathbf{x})$ (where $q(\mathbf{x})$ is normalized, for simplicity).

2. Draw $\mathbf{x}_1 \sim K(\mathbf{x}_1|\mathbf{x}_0)$, where the kernel K leaves invariant density $\bar{\eta}(\mathbf{x}) = \frac{1}{c}\eta(\mathbf{x})$, i.e.,

$$\int_{\mathcal{X}} K(\mathbf{x}'|\mathbf{x})\bar{\eta}(\mathbf{x})d\mathbf{x} = \bar{\eta}(\mathbf{x}'). \tag{107}$$

3. Assign to $\mathbf{x}_1$ the weight

$$\rho(\mathbf{x}_0, \mathbf{x}_1) = \frac{\eta(\mathbf{x}_0)}{q(\mathbf{x}_0)}\frac{\pi(\mathbf{x}_1)}{\eta(\mathbf{x}_1)}. \tag{108}$$

This weight is *proper* in the sense that can be used for building unbiased estimator Z (or other moments $\bar{\pi}(\mathbf{x})$), as described in the Liu's definition [74, Section 14.2], [45, Section 2.5.4]. Indeed, we can write

$$\begin{aligned}
\mathbb{E}[\rho(\mathbf{x}_0, \mathbf{x}_1)] &= \int_{\mathcal{X}}\int_{\mathcal{X}} \rho(\mathbf{x}_0, \mathbf{x}_1)K(\mathbf{x}_1|\mathbf{x}_0)q(\mathbf{x}_0)d\mathbf{x}_0 d\mathbf{x}_1, \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \frac{\eta(\mathbf{x}_0)}{q(\mathbf{x}_0)}\frac{\pi(\mathbf{x}_1)}{\eta(\mathbf{x}_1)}K(\mathbf{x}_1|\mathbf{x}_0)q(\mathbf{x}_0)d\mathbf{x}_0 d\mathbf{x}_1, \\
&= \int_{\mathcal{X}} \frac{\pi(\mathbf{x}_1)}{\eta(\mathbf{x}_1)}\left[\int_{\mathcal{X}} \eta(\mathbf{x}_0)K(\mathbf{x}_1|\mathbf{x}_0)d\mathbf{x}_0\right] d\mathbf{x}_1, \\
&= \int_{\mathcal{X}} \frac{\pi(\mathbf{x}_1)}{c\bar{\eta}(\mathbf{x}_1)}c\bar{\eta}(\mathbf{x}_1)d\mathbf{x}_1 = \int_{\mathcal{X}} \pi(\mathbf{x}_1)d\mathbf{x}_1 = Z.
\end{aligned} \tag{109}$$

Note that if $\eta(\mathbf{x}) \equiv \pi(\mathbf{x})$ then $\rho(\mathbf{x}_1) = \frac{\pi(\mathbf{x}_0)}{q(\mathbf{x}_0)}$, i.e., the IS weights remain unchanged after an MCMC iteration with invariant density $\pi(\mathbf{x})$. Hence, if we repeat the procedure above N times generating $\{\mathbf{x}_0^{(n)}, \mathbf{x}_1^{(n)}\}_{n=1}^N$, we can build the following unbiased estimator of the Z,

$$\widehat{Z} = \frac{1}{N}\sum_{n=1}^{N} \rho(\mathbf{x}_0^{(n)}, \mathbf{x}_1^{(n)}) = \frac{1}{N}\sum_{n=1}^{N} \frac{\eta(\mathbf{x}_0^{(n)})}{q(\mathbf{x}_0^{(n)})}\frac{\pi(\mathbf{x}_1^{(n)})}{\eta(\mathbf{x}_1^{(n)})} \tag{110}$$

In the next section, we extend this idea where different MCMC updates are applied, each one addressing a different invariant density.

4.1.2 Annealed Importance Sampling (An-IS)

In the previous section, we have considered the application of one MCMC kernel $K(\mathbf{x}_1|\mathbf{x}_0)$ (that could be formed by different MCMC steps). Below, we consider the application of several MCMC kernels addressing different target pdfs, and show their consequence in the weighting strategy. We consider again a sequence of tempered versions of the posterior, $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \ldots, \pi_L(\mathbf{x}) \equiv \pi(\mathbf{x})$, where the L-th version, $\pi_L(\mathbf{x})$, coincides with the target function $\pi(\mathbf{x})$. On possibility is to define different scaled version of the target,

$$\pi_i(\mathbf{x}) = [\pi(\mathbf{x})]^{\beta_i} = g(\mathbf{x})^{\beta_i}\ell(\mathbf{y}|\mathbf{x})^{\beta_i}, \quad \text{where} \quad 0 < \beta_1 \leq \beta_2 \leq \ldots \leq \beta_L = 1, \tag{111}$$

or alternatively

$$\pi_i(\mathbf{x}) = g(\mathbf{x})\ell(\mathbf{y}|\mathbf{x})^{\beta_i} \quad \text{where} \quad 0 < \beta_1 \leq \beta_2 \leq \ldots \leq \beta_L = 1. \tag{112}$$

as in path sampling and power posteriors. In any case, smaller β values correspond to flatter distributions.[3] The use of the tempered sequence of target pdfs usually improve the mixing of the algorithm and foster the exploration of the space $\mathcal{X}$. Since only the last function is the true target, $\pi_L(\mathbf{x}) = \pi(\mathbf{x})$, different schemes have been proposed for suitable weighting the final samples.
Let us consider conditional $L-1$ kernels $K_i(\mathbf{z}|\mathbf{x})$ (with $L \geq 2$), representing the probability of different MCMC updates of jumping from the state $\mathbf{x}$ to the state $\mathbf{z}$ (note that each K_i can summarize the application of several MCMC steps), each one leaving invariant a different tempered target, $\bar{\pi}_i(\mathbf{x}) \propto \pi_i(\mathbf{x})$. For one single sample, the Annealed Importance Sampling (An-IS) is given in Table 8.

Table 8: Annealed Importance Sampling (An-IS)

1. Draw N samples $\mathbf{x}_0^{(n)} \sim \bar{\pi}_0(\mathbf{x}) = q(\mathbf{x})$ for $n = 1, ..., N$.
2. For $k = 1, \dots, L-1$:
    (a) Draw N samples $\mathbf{x}_k^{(n)} \sim K_k(\mathbf{x}|\mathbf{x}_{k-1}^{(n)})$, leaving invariant $\bar{\pi}_k(\mathbf{x})$ for $n = 1, ..., N$, i.e., we generate N samples using an MCMC with invariant distribution $\bar{\pi}_k(\mathbf{x})$.

    (b) Compute the weight associated to the sample $\mathbf{x}_k^{(n)}$,

    $$\rho_k^{(n)} = \prod_{i=0}^{k} \frac{\pi_{i+1}(\mathbf{x}_i^{(n)})}{\pi_i(\mathbf{x}_i^{(n)})} = \rho_{k-1}^{(n)} \frac{\pi_{k+1}(\mathbf{x}_k^{(n)})}{\pi_k(\mathbf{x}_k^{(n)})}. \tag{113}$$
3. Return the weighted sample $\{\mathbf{x}_{L-1}^{(n)}, \rho_{L-1}^{(n)}\}_{n=1}^N$. An estimator of the marginal likelihood is $\widehat{Z} = \frac{1}{N}\sum_{n=1}^N \rho_{L-1}^{(n)}$. Combinations of An-IS with path sampling and power posterior methods can be also considered, employing the information of the rest of intermediate densities.

Note that, when $L = 2$, we have $\rho_1^{(n)} = \frac{\pi_1(\mathbf{x}_0^{(n)})}{q(\mathbf{x}_0^{(n)})}\frac{\pi(\mathbf{x}_1^{(n)})}{\pi_1(\mathbf{x}_1^{(n)})}$. If, $\pi_1 = \pi_2 = \ldots = \pi_{L-1} = \eta \neq \pi$, then the weight is $\rho_{L-1} = \frac{\eta(\mathbf{x}_0^{(n)})}{q(\mathbf{x}_0^{(n)})}\frac{\pi(\mathbf{x}_{L-1}^{(n)})}{\eta(\mathbf{x}_{L-1}^{(n)})}$.

The method above can be modified incorporating an additional MCMC transition $\mathbf{x}_L \sim K_L(\mathbf{x}|\mathbf{x}_{L-1})$, which leaves invariant $\bar{\pi}_L(\mathbf{x}) = \bar{\pi}(\mathbf{x})$. However, since $\bar{\pi}_L(\mathbf{x})$ is the true target pdf, as we have seen above the weight remains unchanged (see the case $\bar{\eta}(\mathbf{x}) = \bar{\pi}(\mathbf{x})$ in the previous section). Hence, in this scenario, the output would be $\{\mathbf{x}_L^{(n)}, \rho_L^{(n)}\} = \{\mathbf{x}_L^{(n)}, \rho_{L-1}^{(n)}\}$, i.e., $\rho_L^{(n)} = \rho_{L-1}^{(n)}$. This method has been proposed in [68] but similarly schemes can be found in [13, 29].

Remark. The stepping-stones (SS) sampling method described in Section 3.3 is strictly connected to an Ann-IS scheme *when* the intermediate pdfs $\bar{\pi}_\ell(\mathbf{x})$ are chosen as powered posteriors, i.e.,

[3] Another alternative is to use the so-called *data tempering* [13], for instance, setting $\pi_i(\mathbf{x}) \propto p(\mathbf{x}|y_1, \dots, y_{d+i})$, where $d \geq 1$ and $d + L = D_y$ (recall that $\mathbf{y} = [y_1, \dots, y_{D_y}] \in \mathbb{R}^{D_y}$).

$\bar{\pi}_\ell(\mathbf{x}) \propto \ell(\mathbf{y}|\mathbf{x})^\ell g(\mathbf{x})$. In the SS technique, all the samples, drawn from different powered posteriors, are used within an unique estimator of Z.

Interpretation as Standard IS. For the sake of simplicity, here we consider *reversible* kernels, i.e., each kernel satisfies the detailed balance condition

$$\pi_i(\mathbf{x})K_i(\mathbf{z}|\mathbf{x}) = \pi_i(\mathbf{z})K_i(\mathbf{x}|\mathbf{z}) \quad \text{so that} \quad \frac{K_i(\mathbf{z}|\mathbf{x})}{K_i(\mathbf{x}|\mathbf{z})} = \frac{\pi_i(\mathbf{z})}{\pi_i(\mathbf{x})}. \tag{114}$$

We show that the weighting strategy suggested by An-IS can be interpreted as a standard IS weighting considering the following an extended target density, defined in the extended space $\mathcal{X}^L$,

$$\pi_g(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1}) = \pi(\mathbf{x}_{L-1}) \prod_{k=1}^{L-1} K_k(\mathbf{x}_{k-1}|\mathbf{x}_k). \tag{115}$$

Note that π_g has the true target π as a marginal pdf. Let also consider an extended proposal pdf defined as

$$q_g(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1}) = q(\mathbf{x}_0) \prod_{k=1}^{L-1} K_k(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{116}$$

The standard IS weight of an extended sample $[\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1}]$ in the extended space $\mathcal{X}^L$ is

$$w(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1}) = \frac{\pi_g(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1})}{q_g(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1})} = \frac{\pi(\mathbf{x}_{L-1}) \prod_{k=1}^{L-1} K_k(\mathbf{x}_{k-1}|\mathbf{x}_k)}{q(\mathbf{x}_0) \prod_{k=1}^{L-1} K_k(\mathbf{x}_k|\mathbf{x}_{k-1})}. \tag{117}$$

Replacing the expression $\frac{K_i(\mathbf{z}|\mathbf{x})}{K_i(\mathbf{x}|\mathbf{z})} = \frac{\pi_i(\mathbf{z})}{\pi_i(\mathbf{x})}$ in (117), we obtain the Ann-IS weights

$$\begin{aligned} w(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1}) &= \frac{\pi(\mathbf{x}_{L-1})}{q(\mathbf{x}_0)} \prod_{k=1}^{L-1} \frac{\pi_k(\mathbf{x}_{k-1})}{\pi_k(\mathbf{x}_k)}, & (118) \\ &= \frac{\pi_1(\mathbf{x}_0)}{q(\mathbf{x}_0)} \prod_{k=1}^{L-1} \frac{\pi_{k+1}(\mathbf{x}_k)}{\pi_k(\mathbf{x}_k)} = \prod_{k=0}^{L-1} \frac{\pi_{k+1}(\mathbf{x}_k)}{\pi_k(\mathbf{x}_k)} = \rho_{L-1}, & (119) \end{aligned}$$

where we have used $\pi_L(\mathbf{x}) = \pi(\mathbf{x})$ and just rearranged the numerator.

4.1.3 Generic Sequential Monte Carlo

In this section, we describe a sequential IS scheme which encompasses the previous Ann-IS algorithm as a special case. The method described here uses jointly MCMC transitions and, additionally, resampling steps as well. It is called Sequential Monte Carlo (SMC), since we have a sequence of target pdfs $\pi_k(\mathbf{x})$, $k = 1, \dots, L$ [66]. This sequence of target densities can be defined by a state-space model as in a classical particle filtering framework (truly sequential scenario, where the goal is to track dynamic parameters). Alternatively, we can also consider a static scenario as in the previous sections, i.e., the resulting algorithm is an iterative importance sampler where we consider a sequence of *tempered* densities as in Eqs. (111)-(112), where $\pi_L(\mathbf{x}) = \pi(\mathbf{x})$ [66]. Let us again define an extended proposal density in the domain $\mathcal{X}^k$,

$$\widetilde{q}_k(\mathbf{x}_1, \dots, \mathbf{x}_k) = q_1(\mathbf{x}_1) \prod_{i=2}^{k} F_i(\mathbf{x}_i|\mathbf{x}_{i-1}) : \quad \mathcal{X}^k \to \mathbb{R}, \tag{120}$$

where $q_1(\mathbf{x}_1)$ is a marginal proposal and $F_i(\mathbf{x}_i|\mathbf{x}_{i-1})$ are generic forward transition pdfs, that will be used as partial proposal pdfs. Extending the space from $\mathcal{X}^k$ to $\mathcal{X}^{k+1}$ (increasing its dimension), note that we can write the recursive equation

$$\widetilde{q}_{k+1}(\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{x}_{k+1}) = F_{k+1}(\mathbf{x}_{k+1}|\mathbf{x}_k)\widetilde{q}_k(\mathbf{x}_1, \dots, \mathbf{x}_k) : \quad \mathcal{X}^{k+1} \to \mathbb{R}.$$

The marginal proposal pdfs are

$$\begin{aligned}
q_k(\mathbf{x}_k) &= \int_{\mathcal{X}^{k-1}} \widetilde{q}_k(\mathbf{x}_1, \dots, \mathbf{x}_k) d\mathbf{x}_{1:k-1} \\
&= \int_{\mathcal{X}^{k-1}} q_1(\mathbf{x}_1) \prod_{i=2}^{k} F_i(\mathbf{x}_i|\mathbf{x}_{i-1}) d\mathbf{x}_{1:k-1}, \qquad (121) \\
&= \int_{\mathcal{X}} \left[\int_{\mathcal{X}^{k-2}} q_1(\mathbf{x}_1) \prod_{i=2}^{k} F_i(\mathbf{x}_i|\mathbf{x}_{i-1}) d\mathbf{x}_{1:k-2} \right] F_k(\mathbf{x}_k|\mathbf{x}_{k-1}) d\mathbf{x}_{k-1}, \\
&= \int_{\mathcal{X}} q_{k-1}(\mathbf{x}_{k-1}) F_k(\mathbf{x}_k|\mathbf{x}_{k-1}) d\mathbf{x}_{k-1}, \qquad (122)
\end{aligned}$$

Therefore, we would be interested in computing the *marginal* IS weights, $w_k = \frac{\pi_k(\mathbf{x}_k)}{q_k(\mathbf{x}_k)}$, for each k. However note that, in general, the marginal proposal pdfs $q_k(\mathbf{x}_k)$ cannot be computed and then cannot be evaluated. A suitable alternative approach is described next. Let us consider the extended target pdf defined as

$$\widetilde{\pi}_k(\mathbf{x}_1, \dots, \mathbf{x}_k) = \pi_k(\mathbf{x}_k) \prod_{i=2}^{k} B_{i-1}(\mathbf{x}_{i-1}|\mathbf{x}_i) : \quad \mathcal{X}^k \to \mathbb{R}, \tag{123}$$

$B_{i-1}(\mathbf{x}_{i-1}|\mathbf{x}_i)$ are arbitrary backward transition pdfs. Note that the space of $\{\widetilde{\pi}_k\}$ increases as k grows, and π_k is always a marginal pdf of $\widetilde{\pi}_k$. Moreover, writing the previous equation for $k+1$

$$\widetilde{\pi}_{k+1}(\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{x}_{k+1}) = \pi_{k+1}(\mathbf{x}_{k+1}) \prod_{i=2}^{k+1} B_{i-1}(\mathbf{x}_{i-1}|\mathbf{x}_i),$$

and writing the ratio of both, we get

$$\frac{\widetilde{\pi}_{k+1}(\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{x}_{k+1})}{\widetilde{\pi}_k(\mathbf{x}_1, \dots, \mathbf{x}_k)} = \frac{\pi_{k+1}(\mathbf{x}_{k+1})}{\pi_k(\mathbf{x}_k)} B_k(\mathbf{x}_k|\mathbf{x}_{k+1}). \tag{124}$$

Therefore, the IS weights in the extended space $\mathcal{X}^k$ are

$$\begin{aligned} w_k &= \frac{\widetilde{\pi}_k(\mathbf{x}_1,\dots,\mathbf{x}_k)}{\widetilde{q}_k(\mathbf{x}_1,\dots,\mathbf{x}_k)} && (125)\\ &= \frac{\widetilde{\pi}_{k-1}(\mathbf{x}_1,\dots,\mathbf{x}_{k-1})}{\widetilde{q}_{k-1}(\mathbf{x}_1,\dots,\mathbf{x}_{k-1})}\frac{\frac{\pi_k(\mathbf{x}_k)}{\pi_{k-1}(\mathbf{x}_{k-1})}B_{k-1}(\mathbf{x}_{k-1}|\mathbf{x}_k)}{F_k(\mathbf{x}_k|\mathbf{x}_{k-1})}, && (126)\\ &= w_{k-1}\frac{\pi_k(\mathbf{x}_k)B_{k-1}(\mathbf{x}_{k-1}|\mathbf{x}_k)}{\pi_{k-1}(\mathbf{x}_{k-1})F_k(\mathbf{x}_k|\mathbf{x}_{k-1})}. && (127) \end{aligned}$$

where we have replaced $w_{k-1} = \frac{\widetilde{\pi}_{k-1}(\mathbf{x}_1,\dots,\mathbf{x}_{k-1})}{\widetilde{q}_{k-1}(\mathbf{x}_1,\dots,\mathbf{x}_{k-1})}$. The recursive formula in Eq. (127) is the key expression for several sequential IS techniques. The SMC scheme summarized in Table 9 is a general framework which contains different algorithms as a special cases [66].

Choice of the forward functions. One possible choice is to use independent proposal pdfs, i.e., $F_k(\mathbf{x}_k|\mathbf{x}_{k-1}) = F_k(\mathbf{x}_k)$ or random walk proposal $F_k(\mathbf{x}_k|\mathbf{x}_{k-1})$, where F_k represents standard distributions (e.g., Gaussian or t-Student). An alternative is to choose $F_k(\mathbf{x}_k|\mathbf{x}_{k-1}) = K_k(\mathbf{x}_k|\mathbf{x}_{k-1})$, i.e., an MCMC kernel with invariant pdf $\bar{\pi}_k$.

Choice of backward functions. It is possible to show that the optimal backward transitions $\{B_k\}_{k=1}^L$ are [66]

$$B_{k-1}(\mathbf{x}_{k-1}|\mathbf{x}_k) = \frac{q_{k-1}(\mathbf{x}_{k-1})}{q_k(\mathbf{x}_k)}F_k(\mathbf{x}_k|\mathbf{x}_{k-1}). \qquad (130)$$

This choice reduces the variance of the weights [66]. However, generally, the marginal proposal q_k in Eq. (121) cannot be computed (are not available), other possible $\{B_k\}$ should be considered. For instance, with the choice

$$B_{k-1}(\mathbf{x}_{k-1}|\mathbf{x}_k) = \frac{\pi_k(\mathbf{x}_{k-1})}{\pi_k(\mathbf{x}_k)}F_k(\mathbf{x}_k|\mathbf{x}_{k-1}), \qquad (131)$$

we obtain

$$\begin{aligned} w_k &= w_{k-1}\frac{\pi_k(\mathbf{x}_k)\frac{\pi_k(\mathbf{x}_{k-1})}{\pi_k(\mathbf{x}_k)}F_k(\mathbf{x}_k|\mathbf{x}_{k-1})}{\pi_{k-1}(\mathbf{x}_{k-1})F_k(\mathbf{x}_k|\mathbf{x}_{k-1})} && (132)\\ &= w_{k-1}\frac{\pi_k(\mathbf{x}_{k-1})}{\pi_{k-1}(\mathbf{x}_{k-1})} && (133) \end{aligned}$$

Moreover, if also $F_k(\mathbf{x}_k|\mathbf{x}_{k-1}) = K_k(\mathbf{x}_k|\mathbf{x}_{k-1})$ is an MCMC kernel with invariant $\bar{\pi}_k$, then we come back to An-IS algorithm [68, 13, 29], described in Table 8. Several other methods are contained as special cases of algorithm in Table 9, with specific choice of $\{B_k\}$, $\{K_k\}$ and $\{\pi_k\}$ (e.g., Population Monte Carlo schemes [7]).

Table 9: General Sequential Monte Carlo (SMC)

```
1. Draw $\mathbf{x}_1^{(n)} \sim q_1(\mathbf{x})$, $n = 1, \dots, N$.
2. For $k = 2, \dots, L$ :
    (a) Draw $N$ samples $\mathbf{x}_k^{(n)} \sim F_k(\mathbf{x}|\mathbf{x}_{k-1}^{(n)})$.
    (b) Compute the weights
        $$w_k^{(n)} = w_{k-1}^{(n)} \frac{\pi_k(\mathbf{x}_k^{(n)}) B_{k-1}(\mathbf{x}_{k-1}^{(n)}|\mathbf{x}_k^{(n)})}{\pi_{k-1}(\mathbf{x}_{k-1}^{(n)}) F_k(\mathbf{x}_k|\mathbf{x}_{k-1}^{(n)})}, \quad (128)$$
        $$= w_{k-1}^{(n)} \gamma_k^{(n)}, \qquad , k = 1, \dots, L, \quad (129)$$
        where we set $\gamma_k^{(n)} = \frac{\pi_k(\mathbf{x}_k^{(n)}) B_{k-1}(\mathbf{x}_{k-1}^{(n)}|\mathbf{x}_k^{(n)})}{\pi_{k-1}(\mathbf{x}_{k-1}^{(n)}) F_k(\mathbf{x}_k|\mathbf{x}_{k-1}^{(n)})}$.
    (c) Normalize the weights $\bar{w}_k^{(n)} = \frac{w_k^{(n)}}{\sum_{j=1}^N w_k^{(j)}}$, for $n = 1, ..., N$.
    (d) If $\widehat{ESS} \leq \epsilon N$:
        (with $0 < \epsilon < 1$ and $\widehat{ESS}$ is a effective sample size measure [55], see section 4.1.4)
        i. Resample $N$ times $\{\mathbf{x}_k^{(1)}, \dots, \mathbf{x}_k^{(N)}\}$ according to $\{\bar{w}_k^{(n)}\}_{n=1}^N$, obtaining $\{\bar{\mathbf{x}}_k^{(1)}, \dots, \bar{\mathbf{x}}_k^{(N)}\}$.
        ii. Set $\mathbf{x}_k^{(n)} = \bar{\mathbf{x}}_k^{(n)}$, $\widehat{Z}_k = \frac{1}{N} \sum_{n=1}^N w_k^{(n)}$ and $w_k^{(n)} = \widehat{Z}_k$ for all $n = 1, \dots, N$ [54, 53, 67, 61].
3. Return the cloud of weighted particles and $\widehat{Z} = \widehat{Z}_L = \frac{1}{N} \sum_{n=1}^N w_L^{(n)}$, if a proper weighting of the resampled particles is used (as suggested in the step 2d-ii above). Otherwise, use other estimator $\widehat{Z}_L$ as in Eq. (136) (for further details, see Section 4.1.4).
```

4.1.4 Evidence computation in a sequential framework

The generic algorithm in Table 9 employs also resampling steps. Resampling consists in drawing particles from the current cloud according to the normalized importance weights $\bar{w}_k^{(n)}$, for $n = 1, ..., N$. The resampling steps are applied only in certain iterations taking into account an ESS approximation, such as $\widehat{ESS} = \frac{1}{\sum_{n=1}^N (\bar{w}_k^{(n)})^2}$, or $\widehat{ESS} = \frac{1}{\max_n \bar{w}_k^{(n)}}$ [41, 55]. Generally,If $\frac{1}{N}\widehat{ESS}$ is smaller than a pre-established threshold $\epsilon \in [0, 1]$, all the particles are resampled. Thus, the condition for the adaptive resampling can be expressed as $\widehat{\mathrm{ESS}} < \epsilon N$. When $\epsilon = 1$, the resampling is applied at each iteration [20, 21]. If $\epsilon = 0$, no resampling steps are applied, and we have the SIS method described above. Here, we discuss separately different scenarios $\epsilon = 0$ and $0 < \epsilon \leq 1$ how it is possible to compute sequentially the estimator of the marginal likelihood.

Sequential Importance Sampling (SIS), $\epsilon = 0$. This a specific case of the generic SMC in Table 9, where no resampling steps are applied. Consider the weight recursion in Eq. (128)

$$w_k^{(n)} \;=\; w_{k-1}^{(n)}\gamma_k^{(n)},$$

at the k-th iteration, we have two possible *equivalent* marginal likelihood estimators [54, 53, 61],

$$\widehat{Z}_k^{(1)} = \frac{1}{N}\sum_{n=1}^{N} w_k^{(n)} \tag{134}$$

$$= \frac{1}{N}\sum_{n=1}^{N} w_{k-1}^{(n)}\gamma_k^{(n)} = \frac{1}{N}\sum_{n=1}^{N}\prod_{j=1}^{k}\gamma_j^{(n)}, \tag{135}$$

or

$$\widehat{Z}_k^{(2)} = \prod_{j=1}^{k}\left[\sum_{n=1}^{N}\bar{w}_{j-1}^{(n)}\gamma_j^{(n)}\right]. \tag{136}$$

In SIS, they are completely equivalent, i.e., $\widehat{Z}_k^{(1)} = \widehat{Z}_k^{(2)}$, indeed

$$\widehat{Z}_k^{(2)} = \prod_{j=1}^{k}\left[\sum_{n=1}^{N}\frac{w_{j-1}^{(n)}}{N\widehat{Z}_{j-1}^{(1)}}\gamma_j^{(n)}\right], \tag{137}$$

$$= \prod_{j=1}^{k}\frac{1}{N\widehat{Z}_{j-1}^{(1)}}\left[\sum_{n=1}^{N}w_{j-1}^{(n)}\gamma_j^{(n)}\right], \tag{138}$$

$$= \prod_{j=1}^{k}\frac{1}{N\widehat{Z}_{j-1}^{(1)}}\left[\sum_{n=1}^{N}w_{j}^{(n)}\right], \tag{139}$$

$$= \prod_{j=1}^{k}\frac{\widehat{Z}_{j}^{(1)}}{\widehat{Z}_{j-1}^{(1)}} = \widehat{Z}_k^{(1)}. \tag{140}$$

Therefore, in SIS, we can use both indifferently.

Generic SMC ($0 < \epsilon \leq 1$) with proper weighting after resampling. Let consider that a proper weighting for resampled particles is used [54, 53] as applied in Table 9, i.e., the *unnormalized* weights after resampling are set equal to $\widehat{Z}_k^{(1)} = \frac{1}{N}\sum\limits_{n=1}^{N} w_k^{(n)}$,

$$w_d^{(1)} = w_d^{(2)} = \ldots = w_k^{(N)} = \widehat{Z}_k^{(1)}. \tag{141}$$

Then, it is possible to show that $\widehat{Z}_k^{(1)}$ and $\widehat{Z}_k^{(2)}$ are again equivalent, $\widehat{Z}_k^{(1)} = \widehat{Z}_k^{(2)}$. It can be easily seen for (139),

$$\widehat{Z}_k^{(2)} = \prod_{j=1}^{k}\frac{1}{N\widehat{Z}_{j-1}^{(1)}}\left[\sum_{n=1}^{N}w_{j}^{(n)}\right],$$

$$\widehat{Z}_k^{(2)} = \prod_{j=1}^{k}\frac{1}{N\widehat{Z}_{j-1}^{(1)}}\left[N\widehat{Z}_j^{(1)}\right] = \prod_{j=1}^{k}\frac{\widehat{Z}_{j}^{(1)}}{\widehat{Z}_{j-1}^{(1)}} = \widehat{Z}_k^{(1)}. \tag{142}$$

Note that, we alway have $\bar{w}_k^{(1)} = \bar{w}_k^{(2)} = \ldots = \bar{w}_k^{(N)} = \frac{1}{N}$.

Generic SMC ($0 < \epsilon \leq 1$) *without* **proper weighting after resampling.** In many works regarding particle filtering it is noted that the *unnormalized* weights of the resampled particles,

$$w_k^{(1)} = w_k^{(2)} = \ldots = w_k^{(N)},$$

but a specific value is not assigned (or usually people set to 1). However, also in this case, we always have

$$\bar{w}_k^{(1)} = \bar{w}_k^{(2)} = \ldots = \bar{w}_k^{(N)} = \frac{1}{N}, \tag{143}$$

then $\widehat{Z}_k^{(2)}$ can still be applied. However, $\widehat{Z}_k^{(1)}$, which involves the unnormalized weights, is not more a valid estimator (it loses its statistical meaning) [54, 53].

4.2 Estimation based on Multiple Try MCMC schemes

The Multiple Try Metropolis (MTM) methods are advanced MCMC algorithms which consider different candidates as possible new state of the chain [49, 60, 59]. More specifically, at each iteration different samples are generated and compared by using some proper weights. Then one of them is selected and tested as possible future state. The main advantage of these algorithms is that they foster the exploration of a larger portion of the sample space, decreasing the correlation among the states of the generated chain. Here, we consider the use of importance weights for comparing the different candidates, in order to provide also an estimation of the marginal likelihood [60]. More specifically, we consider the Independent Multiple Try Metropolis type 2 (IMTM-2) scheme [49] with an adaptive proposal pdf. The algorithm is given in Table 10. the mean vector and covariance matrix are adapted using the empirical estimators yielding by all the weighted candidates drawn so far, i.e., $\{\mathbf{z}_{n,\tau}, w_{n,\tau}\}$ for all $n = 1, ..., N$ and $\tau = 1, ..., T$. Two possible estimators of the marginal likelihood can be constructed, one based on standard adaptive importance sampling argument $\widehat{Z}^{(2)}$ [5, 6] and other based on group importance sampling idea provided in [53].

4.3 Layered Adaptive Importance Sampling (LAIS)

The LAIS algorithm consider the use of N parallel (independent or interacting) MCMC chains with invariant pdf $\bar{\pi}(\mathbf{x})$ or a tempered version $\bar{\pi}(\mathbf{x}|\beta)$ [57, 5]. Each MCMC chain can addressed a different tempered version $\bar{\pi}(\mathbf{x}|\beta)$. After T iterations of the N MCMC schemes (upper layer), the resulting NT samples, $\{\boldsymbol{\mu}_{n,t}\}$, for $n = 1, ..., N$ and $t = 1, ..., T$ are used are location parameters of NT proposal densities $q(\mathbf{x}|\boldsymbol{\mu}_{n,t}, \mathbf{C})$. Then, these proposal pdfs are employed within a MIS scheme (lower layer), weighting the generated samples $\mathbf{x}_{n,t}$'s with the generic weight $w_{n,t} = \frac{\pi(\mathbf{x}_{n,t})}{\Phi(\mathbf{x}_{n,t})}$ [24, 22]. The denominator $\Phi(\mathbf{x}_{n,t})$ is a mixture of (all or a subset of) proposal densities which specifies the type of MIS scheme applied [24, 22]. The algorithm, with different possible choices of $\Phi(\mathbf{x}_{n,t})$, is shown in Table 11. The first choice in (148) is the most costly since we have to evaluate all the proposal pdfs in all the generated samples $\mathbf{x}_{n,t}$'s, but provides the best performance in terms of

Table 10: Adaptive Independent Multiple Try Metropolis type 2 (AIMTM-2)

1. Choose the initial parameters $\boldsymbol{\mu}_t$, $\mathbf{C}_t$ of the proposal q, and initial state $\mathbf{x}_0$ and a first estimation of the marginal likelihood $\widehat{Z}_0$.

2. For $t = 1, ..., T$:

   (a) Draw $\mathbf{z}_{1,t}, ..., \mathbf{z}_{N,t} \sim q(\mathbf{z}|\boldsymbol{\mu}_t, \mathbf{C}_t)$.

   (b) Compute the importance weights $w_{n,t} = \frac{\pi(\mathbf{z}_{n,t})}{q(\mathbf{z}_{n,t}|\boldsymbol{\mu}_t, \mathbf{C}_t)}$, for $n = 1, ..., N$.

   (c) Normalize them $\bar{w}_{n,t} = \frac{w_{n,t}}{N\widehat{Z}'}$ where

   $$\widehat{Z}' = \frac{1}{N}\sum_{i=1}^{N} w_{i,t}, \qquad \text{and set} \qquad R_t = \widehat{Z}'. \tag{144}$$

   (d) Resample $\mathbf{x}' \in \{\mathbf{z}_{1,t}, ..., \mathbf{z}_{N,t}\}$ according to $\bar{w}_n$, with $n = 1, ..., N$.

   (e) Set $\mathbf{x}_t = \mathbf{x}'$ and $\widehat{Z}_t = \widehat{Z}'$ with probability

   $$\alpha = \min\left[1, \frac{\widehat{Z}'}{\widehat{Z}_{t-1}}\right] \tag{145}$$

   otherwise set $\mathbf{x}_t = \mathbf{x}_{t-1}$ and $\widehat{Z}_t = \widehat{Z}_{t-1}$.

   (f) Update $\boldsymbol{\mu}_t$, $\mathbf{C}_t$ computing the corresponding empirical estimators using $\{\mathbf{z}_{n,\tau}, w_{n,\tau}\}$ for all $n = 1, ..., N$ and $\tau = 1, ..., T$.

3. Return the chain $\{\mathbf{x}_t\}_{t=1}^{T}$, $\{\widehat{Z}_t\}_{t=1}^{T}$ and $\{R_t\}_{t=1}^{T}$. Two possible estimators of Z can be constructed:

   $$\widehat{Z}^{(1)} = \frac{1}{T}\sum_{t=1}^{T} \widehat{Z}_t, \qquad \widehat{Z}^{(2)} = \frac{1}{T}\sum_{t=1}^{T} R_t. \tag{146}$$

efficiency of the final estimator. The second and third choices are temporal and spatial mixtures, respectively. The last choice corresponds to standard importance weights given in Section 3.

Let assume $\bar{\pi}_n(\mathbf{x}) = \bar{\pi}(\mathbf{x})$ for all n in the upper layer. Considering also standard parallel Metropolis-Hastings chains in the upper layer, the number of posterior evaluations in LAIS is $2NT$. Thus, if only one chain $N = 1$ is employed in the upper layer, the number of posterior evaluations is $2T$.

Special case with recycling samples. The method in [78] can be considered as a special case of LAIS when $N = 1$, and $\{\boldsymbol{\mu}_t = \mathbf{x}_t\}$ i.e., all the samples $\{\mathbf{x}_t\}_{t=1}^{T}$ are generated by the unique MCMC chain with random walk proposal $\varphi(\mathbf{x}|\mathbf{x}_{t-1}) = q(\mathbf{x}|\mathbf{x}_{t-1})$ with invariant density $\bar{\pi}(\mathbf{x})$. In this scenario, the two layers of LAIS are collapsed in a unique layer, so that $\{\boldsymbol{\mu}_t = \mathbf{x}_t\}$. Namely, no additional generation of samples are needed in the lower layer, and the samples generated in

Table 11: Layered Adaptive Importance Sampling (LAIS)

1. Generate NT samples, $\{\boldsymbol{\mu}_{n,t}\}$, using N parallel MCMC chains of length T, each MCMC method using a proposal pdf $\varphi_n(\boldsymbol{\mu}|\boldsymbol{\mu}_{t-1})$, with invariant distributions a power posterior $\bar{\pi}_n(\mathbf{x}) = \bar{\pi}(\mathbf{x}|\beta_n)$ (with $\beta_n > 0$) or a posterior pdf with a smaller number of data.

2. Draw NT samples $\mathbf{x}_{n,t} \sim q_{n,t}(\mathbf{x}|\boldsymbol{\mu}_{n,t}, \mathbf{C})$ where $\boldsymbol{\mu}_{n,t}$ plays the role of the mean, and $\mathbf{C}$ is a covariance matrix.

3. Assign to $\mathbf{x}_{n,t}$ the weights

$$w_{n,t} = \frac{\pi(\mathbf{x}_{n,t})}{\Phi(\mathbf{x}_{n,t})}. \tag{147}$$

   There are different possible choices for $\Phi(\mathbf{x}_{n,t})$, for instance:

$$\begin{aligned}
\Phi(\mathbf{x}_{n,t}) &= \frac{1}{NT}\sum_{k=1}^{T}\sum_{i=1}^{N} q_{n,t}(\mathbf{x}_{n,t}|\boldsymbol{\mu}_{i,k}, \mathbf{C}), && (148)\\
\Phi(\mathbf{x}_{n,t}) &= \frac{1}{T}\sum_{k=1}^{T} q_{n,t}(\mathbf{x}_{n,t}|\boldsymbol{\mu}_{n,k}, \mathbf{C}), && (149)\\
\Phi(\mathbf{x}_{n,t}) &= \frac{1}{N}\sum_{i=1}^{N} q_{n,t}(\mathbf{x}_{n,t}|\boldsymbol{\mu}_{i,t}, \mathbf{C}), && (150)\\
\Phi(\mathbf{x}_{n,t}) &= q_{n,t}(\mathbf{x}_{n,t}|\boldsymbol{\mu}_{i,t}, \mathbf{C}), && (151)
\end{aligned}$$

4. Return all the pairs $\{\mathbf{x}_{n,t}, w_{n,t}\}$, and $\widehat{Z} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{n=1}^{N} w_{n,t}$.

the upper layer (via MCMC) are recycled. Hence, the number of posterior evaluations is only T. The denominator for weights used in [78] is in Eq. (149), i.e., a temporal mixture as in [17]. The resulting estimator is

$$\widehat{Z} = \frac{1}{T}\sum_{t=1}^{T} \frac{\pi(\mathbf{x}_t)}{\frac{1}{T}\sum_{k=1}^{T}\varphi(\mathbf{x}_t|\mathbf{x}_{t-1})}, \quad \{\mathbf{x}_t\}_{t=1}^{T} \sim \bar{\pi}(\mathbf{x}) \text{ (via MCMC with a proposal } \varphi(\cdot|\cdot)).$$

Relationship with KDE method. LAIS can be interpreted as an extension of the KDE method in Section 2, where the KDE function is also employed as a proposal density in the MIS scheme. Namely, the points used in Eq. (18), in LAIS they are drawn from the KDE function using the deterministic mixture procedure [24, 23, 22].

Compressed LAIS (CLAIS). Let us consider the T or N is large (i.e., either large chains or several parallel chains; or both). Since NT is large, the computation of the denominators Eqs. (148)- (149)- (150) can be expensive. A possible solution is to use a partitioning or clustering procedure [52] with $K << NT$ clusters considering the NT samples, and then employ as denominator the function

$$\Phi(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\mathbf{x}|\bar{\boldsymbol{\mu}}_k, \mathbf{C}_k), \tag{152}$$

where $\bar{\boldsymbol{\mu}}_k$ represents the centroid of the k-th cluster, and $\mathbf{C}_k = \boldsymbol{\Sigma}_k + h\mathbf{I}$ with $\boldsymbol{\Sigma}_k$ the empirical covariance matrix of k-th cluster and $h > 0$.

Relationship with power-posteriors methods. In the upper layer of LAIS, we can use non-tempered versions of the posterior, i.e., $\bar{\pi}_n(\mathbf{x}) = \bar{\pi}(\mathbf{x})$ for all n, or tempered versions of the posterior $\bar{\pi}_n(\mathbf{x}) = \bar{\pi}(\mathbf{x}|\beta_n) = \ell(\mathbf{y}|\mathbf{x})^{\beta_n} g(\mathbf{x})$. However, unlike in power posterior methods these samples are employed only as location parameters $\boldsymbol{\mu}_{n,t}$ of the proposal pdfs $q_{n,t}(\mathbf{x}|\boldsymbol{\mu}_{n,t}, \mathbf{C})$, and they are not included in the final estimators. Combining the power-posteriors idea and the approach in [78], we could recycle $\mathbf{x}_{n,t} = \boldsymbol{\mu}_{n,t}$ and use $q_{n,t}(\mathbf{x}|\boldsymbol{\mu}_{n,t}) = \varphi_{n,t}(\mathbf{x}|\boldsymbol{\mu}_{n,t})$ where we denote as $\varphi_{n,t}$ the proposal pdfs employed in the MCMC chains. Another difference is that in LAIS the use of an "anti-tempered" power posterior with $\beta_n > 1$ is allowed and can be shown that is beneficial for the performance of the estimators (after that the chains reach a good mixing) [56]. More generally, one can consider a time-varying $\beta_{n,t}$ (where t is the iteration of the n-th chain). In the first iterations, one could use $\beta_{n,t} < 1$ for fostering the exploration of the state space and helping the mixing of the chain. Then, in the last iterations, one could use $\beta_{n,t} > 1$ which increases the efficiency of the resulting IS estimators [56].

5 Vertical likelihood representations

5.1 Lebesgue representations of the marginal likelihood

5.1.1 First one-dimensional representation

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

$$\begin{aligned}
Z &= \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})d\mathbf{x} &(153)\\
&= \int_{\mathcal{X}} g(\mathbf{x})d\mathbf{x} \int_0^{\ell(\mathbf{y}|\mathbf{x})} d\lambda \quad \text{(extended space representation)} &(154)\\
&= \int_{\mathcal{X}} g(\mathbf{x})d\mathbf{x} \int_0^{\infty} \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\mathbf{x})\}d\lambda &(155)
\end{aligned}$$

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

$$\begin{aligned} Z &= \int_0^\infty d\lambda \int_{\mathcal{X}} g(\mathbf{x}) \mathbb{I}\{0 < \lambda < \ell(\mathbf{y}|\mathbf{x})\} d\mathbf{x} && (156) \\ &= \int_0^\infty d\lambda \int_{\ell(\mathbf{y}|\mathbf{x})>\lambda} g(\mathbf{x}) d\mathbf{x} && (157) \\ &= \int_0^\infty Z(\lambda) d\lambda = \int_0^{\sup \ell(\mathbf{y}|\mathbf{x})} Z(\lambda) d\lambda, && (158) \end{aligned}$$

where we have set

$$Z(\lambda) = \int_{\ell(\mathbf{y}|\mathbf{x})>\lambda} g(\mathbf{x}) d\mathbf{x}. \tag{159}$$

In Eq. (158), we have also assumed that $\ell(\mathbf{y}|\mathbf{x})$ is bounded so the limit of integration is $\sup \ell(\mathbf{y}|\mathbf{x})$.

5.1.2 The survival function $Z(\lambda)$ and related sampling procedures

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

$$Z(\lambda) = \mathbb{P}\left(\lambda < \ell(\mathbf{y}|\mathbf{X})\right), \qquad \text{where } \mathbf{X} \sim g(\mathbf{x}). \tag{160}$$

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

$$F(\lambda) = 1 - Z(\lambda) = \mathbb{P}\left(\ell(\mathbf{y}|\mathbf{X}) < \lambda\right) = \mathbb{P}\left(\Lambda < \lambda\right), \tag{161}$$

is the cumulative distribution of the random variable $\Lambda = \ell(\mathbf{y}|\mathbf{X})$ with $\mathbf{X} \sim g(\mathbf{x})$ [58, 74].

Sampling according to $\boldsymbol{F(\lambda) = 1 - Z(\lambda)}$. Since $\Lambda = \ell(\mathbf{y}|\mathbf{X})$ with $\mathbf{X} \sim g(\mathbf{x})$, the following procedure generates samples λ_n from $\frac{dF(\lambda)}{d\lambda}$:

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

Recalling the inversion method [58, Chapter 2], note also that the corresponding values

$$b_n = F(\lambda_n) \sim \mathcal{U}([0, 1]), \tag{162}$$

i.e., they are uniformly distributed in $[0, 1]$. Since $Z(\lambda) = 1 - F(\lambda)$, and since $V = 1 - U$ is also uniformly distributed $\mathcal{U}([0, 1])$ if $U \sim \mathcal{U}([0, 1])$, then

$$a_n = Z(\lambda_n) \sim \mathcal{U}([0, 1]). \tag{163}$$

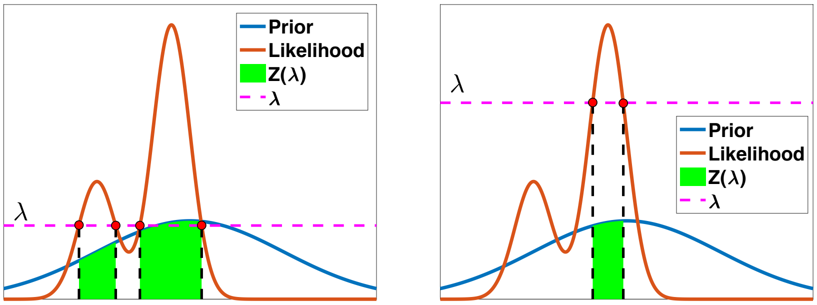
In summary, finally we have that

$$\text{if } \mathbf{x}_n \sim g(\mathbf{x}), \text{ and } \lambda_n = \ell(\mathbf{y}|\mathbf{x}_n) \sim F(\lambda) \quad \text{then} \quad a_n = Z(\lambda_n) \sim \mathcal{U}([0,1]). \tag{164}$$

The truncated prior pdf. Note that $Z(\lambda)$ is also the normalizing constant of the following truncated prior pdf

$$g(\mathbf{x}|\lambda) = \frac{1}{Z(\lambda)} \mathbb{I}\{\ell(\mathbf{y}|\mathbf{x}) > \lambda\} g(\mathbf{x}), \tag{165}$$

where $g(\mathbf{x}|0) = g(\mathbf{x})$ and $g(\mathbf{x}|\lambda)$ for $\lambda > 0$. Two graphical examples of $g(\mathbf{x}|\lambda)$ and $Z(\lambda)$ are given in Figure 2.



(a) (b)

Figure 2: Two examples of the area below the truncated prior $g(\mathbf{x}|\lambda)$, i.e., the function $Z(\lambda)$. Note that in figure (b) the value of λ is greater than in figure (a), so that the area $Z(\lambda)$ decreases. If λ is bigger than the maximum of the likelihood function then $Z(\lambda) = 0$.

Sampling from $g(\mathbf{x}|\boldsymbol{\lambda})$ and $F(\boldsymbol{\lambda}|\boldsymbol{\lambda}_0)$. Given a fixed value $\lambda_0 \geq 0$, in order to generate samples from $g(\mathbf{x}|\lambda_0)$ one alternative is to use an MCMC procedure. However, in this case, the following acceptance-rejection procedure can be also employed [58]:

1. For $n = 1, ..., N$:
   (a) Draw $\mathbf{x}' \sim g(\mathbf{x})$.
   (b) if $\ell(\mathbf{y}|\mathbf{x}') > \lambda_0$ then set $\mathbf{x}_n = \mathbf{x}'$ and $\lambda_n = \ell(\mathbf{y}|\mathbf{x}')$.
   (c) if $\ell(\mathbf{y}|\mathbf{x}') \leq \lambda_0$, then reject $\mathbf{x}'$ and repeat from step 1(a).
2. Return $\{\mathbf{x}_n\}_{n=1}^N$ and $\{\lambda_n\}_{n=1}^N$.

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

$$F(\lambda|\lambda_0) = \frac{F(\lambda) - F(\lambda_0)}{1 - F(\lambda_0)}, \qquad \text{with } \lambda \geq \lambda_0, \tag{166}$$

i.e., we can write $\lambda_n \sim F(\lambda|\lambda_0)$.

Distribution of $a_n = Z(\lambda_n)$ if $\lambda_n \sim F(\lambda|\lambda_0)$. Considering the values $\lambda_n = \ell(\mathbf{y}|\mathbf{x}_n)$ where $\mathbf{x}_n \sim g(\mathbf{x}|\lambda_0)$, then $\lambda_n \sim F(\lambda|\lambda_0)$. Therefore, considering the values $a_0 = Z(\lambda_0) \leq 1$ and $a_n = Z(\lambda_n)$, with a similar argument used above in Eqs. (163)-(164) we can write

$$a_n \sim \mathcal{U}([0, a_0]), \tag{167}$$

$$\widetilde{a}_n = \frac{a_n}{a_0} \sim \mathcal{U}([0, 1]), \quad \forall n = 1, ..., N. \tag{168}$$

In summary, with $a_0 = Z(\lambda_0)$, we have that

$$\text{if } \mathbf{x}_n \sim g(\mathbf{x}|\lambda_0) \text{ and } \lambda_n = \ell(\mathbf{y}|\mathbf{x}_n) \sim F(\lambda|\lambda_0), \quad \text{then} \quad Z(\lambda_n) \sim \mathcal{U}([0, a_0]). \tag{169}$$

and the ratio $\widetilde{a}_n = \frac{a_n}{a_0} \sim \mathcal{U}([0, 1])$.

Distributions $\widetilde{a}_{\texttt{max}}$. Let us consider $\lambda_1, ..., \lambda_n \sim F(\lambda|\lambda_0)$ and the minimum and maximum values

$$\lambda_{\texttt{min}} = \min_n \lambda_n, \quad a_{\texttt{max}} = Z(\lambda_{\texttt{min}}), \quad \text{and} \quad \widetilde{a}_{\texttt{max}} = \frac{a_{\texttt{max}}}{a_0} = \frac{Z(\lambda_{\texttt{min}})}{Z(\lambda_0)}. \tag{170}$$

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

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

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

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

is distributed according to a Beta distribution $\mathcal{B}(N, 1)$, i.e., $F_{\texttt{max}}(\widetilde{a}) = \widetilde{a}^N$ and density $f_{\texttt{max}}(\widetilde{a}) = \frac{dF_{\max}(\widetilde{a})}{d\widetilde{a}} = N\widetilde{a}^{N-1}$ [58, Section 2.3.6]. In summary, we have

$$\widetilde{a}_{\texttt{max}} = \frac{Z(\lambda_{\texttt{min}})}{Z(\lambda_0)} \sim \mathcal{B}(N, 1), \text{ where } \lambda_{\texttt{min}} = \min_n \lambda_n, \quad \text{and} \quad \lambda_n \sim F(\lambda|\lambda_0). \tag{171}$$

This result is important for deriving the standard version of the *nested sampling* method, described in the next section.

5.1.3 Second one-dimensional representation

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

$$\Psi(a) = Z^{-1}(a) = \sup\{\lambda : Z(\lambda) > a\}, \tag{172}$$

is also non-increasing. Note that $Z(\lambda) > a$ if and only if $\lambda < \Psi(a)$. Then, we can write

$$\begin{aligned}
Z &= \int_0^{\infty} Z(\lambda) d\lambda && (173)\\
&= \int_0^{\infty} d\lambda \int_0^1 \mathbb{I}\{a < Z(\lambda)\} da \qquad \text{(again the extended space "trick")} && (174)\\
&= \int_0^1 da \int_0^{\infty} \mathbb{I}\{u < Z(\lambda)\} d\lambda \qquad \text{(swicthing the integration order)} && (175)\\
&= \int_0^1 da \int_0^{\infty} \mathbb{I}\{\lambda < \Psi(a)\} d\lambda \qquad (\text{using } Z(\lambda) > a \iff \lambda < \Psi(a)) && (176)\\
&= \int_0^1 \Psi(a) da. && (177)
\end{aligned}$$

5.1.4 Summary of the one-dimensional representations

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

$$Z = \int_0^{\sup \ell(\mathbf{y}|\mathbf{x})} Z(\lambda) d\lambda = \int_0^1 \Psi(a) da. \tag{178}$$

Now that we have expressed the quantity Z as an integral of a function over $\mathbb{R}$, we could think of applying simple quadrature: choose a grid of points in $[0, \sup \ell(\mathbf{y}|\mathbf{x})]$ ($\lambda_i > \lambda_{i-1}$) or in $[0, 1]$ ($a_i > a_{i-1}$), evaluate $Z(\lambda)$ or $\Psi(a)$ and use the quadrature formulas

$$\widehat{Z} = \sum_{i=1}^{I} (\lambda_i - \lambda_{i-1}) Z(\lambda_i), \ \text{ or} \tag{179}$$

$$\widehat{Z} = \sum_{i=1}^{I} (a_i - a_{i-1}) \Psi(a_i). \tag{180}$$

However, this simple approach is not desirable since (i) the functions $Z(\lambda)$ and $\Psi(a)$ are intractable in most cases and (ii) they change much more rapidly over their domains than does $\pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})$, hence the quadrature approximation can have very bad performance, unless the grid of points is chosen with extreme care. Table 12 summarizes the one-dimensional expression for $\log Z$ and Z contained in this work. Clearly, in all of them, the integrand function depends, explicitly or implicitly, on the variable $\mathbf{x}$.

5.2 Nested Sampling

Nested sampling is a technique for estimating the marginal likelihood that exploits the second identity in (178) [80, 14, 70]. Nested Sampling estimates Z by a quadrature using nodes (in *decreasing* order),

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

Table 12: One-dimensional expression for $\log Z$ and Z. Note that, in all cases, the integrand function contains the dependence on $\mathbf{x}$.

Method	Expression	Equations
power-posteriors	$\log Z = \int_0^1 \mathbb{E}_{\widetilde{\pi}(\mathbf{x}\mid\beta)}\left[\log \ell(\mathbf{y}\mid\mathbf{x})\right] d\beta$	(95)
vertical representation-1	$Z = \int_0^{\sup \ell(\mathbf{y}\mid\mathbf{x})} Z(\lambda) d\lambda$	(158)-(159)
vertical representation-2	$Z = \int_0^1 \Psi(a) da$	(177)

and the quadrature formula

$$\widehat{Z} = \sum_{i=1}^{I-1} (a_{\mathtt{max}}^{(i-1)} - a_{\mathtt{max}}^{(i)}) \Psi(a_{\mathtt{max}}^{(i)}) = \sum_{i=1}^{I-1} (a_{\mathtt{max}}^{(i-1)} - a_{\mathtt{max}}^{(i)}) \lambda_{\mathtt{min}}^{(i)}, \tag{181}$$

with $a_{\mathtt{max}}^{(0)} = 1$. We have to specify the grid points $a_{\mathtt{max}}^{(i)}$'s (possibly well-located, with a suitable strategy) and the corresponding values $\lambda_{\mathtt{min}}^{(i)} = \Psi(a_{\mathtt{max}}^{(i)})$. Recall that the function $\Psi(a)$, and its inverse $a = \Psi^{-1}(\lambda) = Z(\lambda)$, are generally intractable, so that it is not even possible to evaluate $\Psi(a)$ at a grid of chosen $a_{\mathtt{max}}^{(i)}$'s. The nested sampling algorithm works in the other way around: it suitably selects the ordinates $\lambda_{\mathtt{min}}^{(i)}$'s and find some approximations $\widehat{a}_i$'s of the corresponding values $a_{\mathtt{max}}^{(i)} = Z(\lambda_{\mathtt{min}}^{(i)})$. This is possible since the distribution of $a_{\mathtt{max}}^{(i)}$ is known.

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

Nested sampling employs an iterative procedure in order to generate an *increasing* sequence of likelihood ordinates $\lambda_{\mathtt{min}}^{(i)}$, $i = 1, ..., I$, such that

$$\lambda_{\mathtt{min}}^{(1)} < \lambda_{\mathtt{min}}^{(2)} < \lambda_{\mathtt{min}}^{(3)} .... < \lambda_{\mathtt{min}}^{(I)}. \tag{182}$$

The details of the algorithm is given in Table 13 and it is based on the sampling of the truncated prior pdf $g(\mathbf{x}|\lambda_{\mathtt{min}}^{(i-1)})$ (see Section 5.1.2), where i denotes the iteration index. The nested sampling procedure is explained below:

- At the first iteration ($i = 1$), we set $\lambda_{\mathtt{min}}^{(0)} = 0$ and $a_{\mathtt{max}}^{(0)} = Z(\lambda_{\mathtt{min}}^{(0)}) = 1$. Then, N samples are drawn from the prior $\mathbf{x}_n \sim g(\mathbf{x}|\lambda_{\mathtt{min}}^{(0)}) = g(\mathbf{x})$ obtaining a cloud $\mathcal{P} = \{\mathbf{x}_n\}_{n=1}^N$ and then set $\lambda_n = \ell(\mathbf{y}|\mathbf{x}_n)$, i.e., $\{\lambda_n\}_{n=1}^N \sim F(\lambda)$ as shown in Section 5.1.2. Thus, the first ordinate is chosen as
$$\lambda_{\mathtt{min}}^{(1)} = \min_n \lambda_n = \min_n \ell(\mathbf{y}|\mathbf{x}_n) = \min_{\mathbf{x}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P}).$$
Since $\{\lambda_n\}_{n=1}^N \sim F(\lambda)$, using the result in Eq. (171), we have that
$$\widetilde{a}_{\mathtt{max}}^{(1)} = \frac{a_{\mathtt{max}}^{(1)}}{a_{\mathtt{max}}^{(0)}} = \frac{Z(\lambda_{\mathtt{min}}^{(1)})}{Z(\lambda_{\mathtt{min}}^{(0)})} \sim \mathcal{B}(N, 1).$$

Since $a_{\texttt{max}}^{(0)} = Z(\lambda_{\texttt{min}}^{(0)}) = 1$, then $\widetilde{a}_{\texttt{max}}^{(1)} = a_{\texttt{max}}^{(1)} \sim \mathcal{B}(N, 1)$. The corresponding $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$ is also removed from $\mathcal{P}$, i.e., $\mathcal{P} = \mathcal{P}\backslash\{\mathbf{x}^*\}$ (now $|\mathcal{P}| = N - 1$).

- At a generic i-th iteration ($i \geq 2$), a unique additional sample $\mathbf{x}'$ is drawn from the truncated prior $g(\mathbf{x}|\lambda_{\texttt{min}}^{(i-1)})$ and add to the current cloud of samples, i.e., $\mathcal{P} = \mathcal{P} \cup \mathbf{x}'$ (now again $|\mathcal{P}| = N$). First of all, note that the value $\lambda' = \lambda_n = \ell(\mathbf{y}|\mathbf{x}')$ is distributed as $F(\lambda|\lambda_{\texttt{min}}^{(i-1)})$ (see Section 5.1.2). More precisely, note that all the N ordinate values

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

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

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

Moreover, since $\lambda_{\texttt{min}}^{(i)}$ is the minimum value of $\{\lambda_1, ..., \lambda\} \sim F(\lambda|\lambda_{\texttt{min}}^{(i-1)})$, in Section 5.1.2 we have seen that

$$\widetilde{a}_{\texttt{max}}^{(i)} = \frac{a_{\texttt{max}}^{(i)}}{a_{\texttt{max}}^{(i-1)}} = \frac{Z(\lambda_{\texttt{min}}^{(i)})}{Z(\lambda_{\texttt{min}}^{(i-1)})} \sim \mathcal{B}(N, 1), \tag{183}$$

where we have used Eq. (171). We remove again the corresponding sample $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$, i.e., we set $\mathcal{P} = \mathcal{P} \cup \mathbf{x}'$ and the procedure is repeated. Note that we have found the recursion

$$a_{\texttt{max}}^{(i)} = \widetilde{a}_{\texttt{max}}^{(i)} a_{\texttt{max}}^{(i-1)}, \tag{184}$$

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

- A possible idea for approximating the random value $\widetilde{a}_{\texttt{max}}^{(i)}$ is to replace it with the expected value of Beta distribution $\mathcal{B}(N, 1)$, i.e.,

$$\widetilde{a}_{\texttt{max}}^{(i)} \approx \widehat{a}_1 = \frac{N}{N+1} \approx \exp\left(-\frac{1}{N}\right). \tag{185}$$

where $\mathbb{E}[\mathcal{B}(N, 1)] = \frac{N}{N+1}$, and $\exp\left(-\frac{1}{N}\right)$ becomes a very good approximation as N grows. In that case, the recursion above becomes

$$a_{\texttt{max}}^{(i)} \approx \exp\left(-\frac{1}{N}\right) a_{\texttt{max}}^{(i-1)} = \exp\left(-\frac{i}{N}\right). \tag{186}$$

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

The intuition behind the iterative approach above is to accumulate more ordinates λ_i close to the $\sup \ell(\mathbf{y}|\mathbf{x})$. They are also more dense around $\sup \ell(\mathbf{y}|\mathbf{x})$. Moreover, using this scheme, we can provide an approximation $\widehat{a}_i$ of $a_{\texttt{max}}^{(i)}$ since we know the distribution of $\widetilde{a}_{\texttt{max}}^{(i)}$.

Table 13: The standard Nested Sampling procedure.

1. Choose N and set $\widehat{a}_0 = 1$.
2. Draw $\{\mathbf{x}_n\}_{n=1}^N \sim g(\mathbf{x})$ and define the set $\mathcal{P} = \{\mathbf{x}_n\}_{n=1}^N$. Let us also define the notation
$$\ell(\mathbf{y}|\mathcal{P}) = \{\lambda_n = \ell(\mathbf{y}|\mathbf{x}_n) \ \text{ for all } \ \mathbf{x}_n \in \mathcal{P}\}, \tag{187}$$
3. Set $\lambda_{\texttt{min}}^{(1)} = \min_{\mathbf{x}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$ and $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$.
4. Set $\mathcal{P} = \mathcal{P}\backslash\{\mathbf{x}^*\}$, i.e., eliminate $\mathbf{x}^*$ from $\mathcal{P}$.
5. Find an approximation $\widehat{a}_1$ of $a_{\texttt{max}}^{(1)} = Z(\lambda_{\texttt{min}}^{(1)})$. One usual choice is $\widehat{a}_1 = \exp\left(-\frac{1}{N}\right)$.
6. For $i = 2, .., I$:
   (a) Draw $\mathbf{x}' \sim g(\mathbf{x}|\lambda_{\texttt{min}}^{(i-1)})$ and add to the current cloud of samples, i.e., $\mathcal{P} = \mathcal{P} \cup \mathbf{x}'$.
   (b) Set $\lambda_{\texttt{min}}^{(i)} = \min_{\mathbf{x}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$ and $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{P}} \ell(\mathbf{y}|\mathcal{P})$.
   (c) Set $\mathcal{P} = \mathcal{P}\backslash\{\mathbf{x}^*\}$.
   (d) Find an approximation $\widehat{a}_i$ of $a_{\texttt{max}}^{(i)} = Z(\lambda_{\texttt{min}}^{(i)})$. One usual choice is
   $$\widehat{a}_i = \exp\left(-\frac{i}{N}\right), \tag{188}$$
   The rationale behind this choice is explained in the sections above.
7. Return
$$\widehat{Z} = \sum_{i=1}^{I} (\widehat{a}_{i-1} - \widehat{a}_i)\lambda_{\texttt{min}}^{(i)} = \sum_{i=1}^{I} (e^{-\frac{i-1}{N}} - e^{-\frac{i}{N}})\lambda_{\texttt{min}}^{(i)}. \tag{189}$$

5.2.2 Further considerations

Perhaps, the most critical task of the nested sampling implementation consists in drawing from the truncated priors. For this purpose, one can use a rejection sampling or an MCMC scheme. In the first case, we drawn from the prior and then accept only the samples $\mathbf{x}'$ such that $\ell(\mathbf{y}|\mathbf{x}') > \lambda$. However, as λ grows, its performance deteriorates since the acceptance probability gets smaller and smaller. The MCMC algorithms could also have poor performance due to the sample correlation, specially when the support of the constrained prior is formed by disjoint regions or distant modes. Moreover, in the derivation of the standard nested sampling method we have considered different approximations. First of all, for each likelihood value λ_i, its corresponding $a_i = \Psi^{-1}(\lambda_i)$ is approximated by taking the expected value of a Beta random variable. Then this expected value is again approximated with an exponential function in Eq. (185). This step could be avoided, keeping directly $\frac{N}{N+1}$. The simplicity of the final formula $\widehat{a}_i = \exp\left(-\frac{i}{N}\right)$ is perhaps the reason of using the approximation $\frac{N}{N+1} \approx \exp\left(-\frac{1}{N}\right)$. A further approximation $\mathbb{E}[a_{\texttt{max}}^{(i)}] \approx \mathbb{E}[\widetilde{a}_{\texttt{max}}^{(i)}]\mathbb{E}[a_{\texttt{max}}^{(i-1)}]$ is also applied. Additionally, if an MCMC method is run for sampling from the constrained prior, also the likelihood values λ_i are in some sense approximated due to the possible burn-in period of the chain.

5.3 Generalized Importance Sampling based on vertical representations

Let us recall the two possible IS estimators with proposal density $\bar{q}(\mathbf{x})$,

$$\widehat{Z} = \frac{1}{N}\sum_{n=1}^{N} \rho_n \ell(\mathbf{y}|\mathbf{x}_n), \quad \text{and} \quad \widehat{Z} = \sum_{n=1}^{N} \bar{\rho}_n \ell(\mathbf{y}|\mathbf{x}_n), \ \ \{\mathbf{x}_n\}_{n=1}^{N} \sim \bar{q}(\mathbf{x}), \tag{190}$$

where $\rho_n = \frac{g(\mathbf{x}_n)}{\bar{q}(\mathbf{x}_n)}$ and $\bar{\rho}_n = \frac{\rho_n}{\sum_{n=1}^{N} \rho_n}$. In [70], the authors consider the use of the following proposal pdf

$$\bar{q}(\mathbf{x}) = \bar{\pi}_w(\mathbf{x}) = \frac{g(\mathbf{x})W(\ell(\mathbf{y}|\mathbf{x}))}{Z_w}, \tag{191}$$

where the function $W(\lambda) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is defined by the user. Using $\bar{q}(\mathbf{x}) = \bar{\pi}_w(\mathbf{x})$ leads to the weights of the form

$$\rho_n = \frac{g(\mathbf{x}_n)}{\bar{\pi}_w(\mathbf{x}_n)} = \frac{1}{W(\ell(\mathbf{y}|\mathbf{x}_n))}, \quad \mathbf{x}_n \sim \bar{\pi}_w(\mathbf{x}). \tag{192}$$

Note that choosing $W(\lambda) = \lambda$ we have $W(\ell(\mathbf{y}|\mathbf{x})) = \ell(\mathbf{y}|\mathbf{x})$, and $\bar{q}(\mathbf{x}) = \bar{\pi}(\mathbf{x})$, recovering the harmonic mean estimator. With $W(\lambda) = \lambda^\beta$, we have $W(\ell(\mathbf{y}|\mathbf{x})) = \ell(\mathbf{y}|\mathbf{x})^\beta$ and $\bar{q}(\mathbf{x}) = \frac{g(\mathbf{x})\ell(\mathbf{y}|\mathbf{x})^\beta}{Z(\beta)}$, recovering the method in Section 3.3 that uses a power posterior as a proposal pdf. Note that also nested sampling can be included in Eqs. (190), i.e., $\widehat{Z} = \frac{1}{N}\sum_{i=1}^{I} \rho_n \ell(\mathbf{y}|\mathbf{x}_n)$, where $\rho_i = N(e^{-\frac{n-1}{N}} - e^{-\frac{n}{N}})$, that is, a weighted sum of likelihood values $\lambda_n = \ell(\mathbf{y}|\mathbf{x}_n)$.

6 Bayes factors with improper priors

So far we have considered proper priors, i.e., $\int_{\mathcal{X}} g(\mathbf{x})d\mathbf{x} = 1 < \infty$. The use of improper priors is common in Bayesian inference to represent weak prior information. Consider $g(\mathbf{x}) \propto h(\mathbf{x})$ where $h(\mathbf{x})$ is a function whose integral over the state space does not converge, $\int_{\mathcal{X}} g(\mathbf{x})d\mathbf{x} = \int_{\mathcal{X}} h(\mathbf{x})d\mathbf{x} = \infty$. In that case, $g(\mathbf{x})$ is not completely specified. Indeed, we can have different definitions $g(\mathbf{x}) = ch(\mathbf{x})$ where $c > 0$ is the "normalizing" constant, not uniquely determinate since c formally does not exist. Regarding the parameter inference and posterior definition, the use of improper priors poses no problems as long as $\int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})h(\mathbf{x})d\mathbf{x} < \infty$, indeed

$$\bar{\pi}(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x}) = \frac{\ell(\mathbf{y}|\mathbf{x})ch(\mathbf{x})}{\int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})ch(\mathbf{x})d\mathbf{x}} = \frac{\ell(\mathbf{y}|\mathbf{x})h(\mathbf{x})}{\int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})h(\mathbf{x})d\mathbf{x}}, \tag{193}$$

$$= \frac{1}{\widetilde{Z}}\ell(\mathbf{y}|\mathbf{x})h(\mathbf{x}) \tag{194}$$

where $Z = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})d\mathbf{x}$, $\widetilde{Z} = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})h(\mathbf{x})d\mathbf{x}$ and $Z = c\widetilde{Z}$. Note that the unspecified constant $c > 0$ is canceled out. However, the issue is not solved when we compare different models. For instance, the Bayes factors depend on the undetermined constants $c_1, c_2 > 0$ [81],

$$\text{BF}(\mathbf{y}) = \frac{c_1}{c_2}\frac{\int_{\mathcal{X}_1} \ell_1(\mathbf{y}|\mathbf{x}_1)h_1(\mathbf{x}_1)d\mathbf{x}_1}{\int_{\mathcal{X}_2} \ell_2(\mathbf{y}|\mathbf{x}_2)h_2(\mathbf{x}_2)d\mathbf{x}_2} = \frac{Z_1}{Z_2} = \frac{c_1\widetilde{Z}_1}{c_2\widetilde{Z}_2}, \tag{195}$$

so that different choices of c_1, c_2 provide different preferable models. There exists various approaches for dealing with this issue. Below we describe some relevant ones.

Partial Bayes Factors. The idea behind the partial Bayes factors consists of using a subset of data to build proper priors and, jointly with the remaining data, they are used to calculate the Bayes factors. The method starts by dividing the data in two subsets, $\mathbf{y} = (\mathbf{y}_{\text{train}}, \mathbf{y}_{\text{test}})$. The first part $\mathbf{y}_{\text{train}}$ will be used to obtain partial posterior distributions

$$\bar{g}_m(\mathbf{x}_m|\mathbf{y}_{\text{train}}) = \frac{c_m}{Z_{\text{train}}^{(m)}}\ell_m(\mathbf{y}_{\text{train}}|\mathbf{x}_m)h_m(\mathbf{x}_m), \tag{196}$$

using the improper priors. Note that

$$Z_{\text{train}}^{(m)} = c_m \int_{\mathcal{X}_m} \ell_m(\mathbf{y}_{\text{train}}|\mathbf{x}_m)h_m(\mathbf{x}_m)d\mathbf{x}_m.$$

Then, these posteriors can be employed as prior distributions. The complete posterior of m-th model is

$$\bar{\pi}_m(\mathbf{x}|\mathbf{y}) = \bar{\pi}_m(\mathbf{x}|\mathbf{y}_{\text{test}}, \mathbf{y}_{\text{train}}) = \frac{1}{Z_m}\ell_m(\mathbf{y}|\mathbf{x}_m)h_m(\mathbf{x}_m).$$

Considering the conditional likelihood $\ell_m(\mathbf{y}_{\text{test}}|\mathbf{x}_m, \mathbf{y}_{\text{train}})$ of the remaining data $\mathbf{y}_{\text{test}}$, so that we can express the complete posterior as

$$\bar{\pi}_m(\mathbf{x}|\mathbf{y}) = \frac{1}{Z_{\text{test}|\text{train}}^{(m)}}\ell_m(\mathbf{y}_{\text{test}}|\mathbf{x}_m, \mathbf{y}_{\text{train}})\bar{g}_m(\mathbf{x}_m|\mathbf{y}_{\text{train}}), \tag{197}$$

where $\bar{g}_m(\mathbf{x}_m|\mathbf{y}_{\text{train}})$ plays the role of a prior pdf. Replacing the expression of $\bar{g}_m(\mathbf{x}_m|\mathbf{y}_{\text{train}})$, we finally have

$$\bar{\pi}_m(\mathbf{x}|\mathbf{y}) = \frac{c_m}{Z_{\text{test}|\text{train}}^{(m)}Z_{\text{train}}^{(m)}}\ell_m(\mathbf{y}_{\text{test}}|\mathbf{x}_m, \mathbf{y}_{\text{train}})\ell_m(\mathbf{y}_{\text{train}}|\mathbf{x}_m)h_m(\mathbf{x}_m) \tag{198}$$

where $Z_m = Z^{(m)}_{\text{test}|\text{train}} Z^{(m)}_{\text{train}}$, hence

$$Z^{(m)}_{\text{test}|\text{train}} = \frac{Z_m}{Z^{(m)}_{\text{train}}} = \int_{\mathcal{X}_m} \ell_m(\mathbf{y}_{\text{test}}|\mathbf{x}_m, \mathbf{y}_{\text{train}}) \bar{g}_m(\mathbf{x}_m|\mathbf{y}_{\text{train}}) d\mathbf{x}_m. \tag{199}$$

Therefore, considering the partial posteriors $\bar{g}_m(\mathbf{x}_m|\mathbf{y}_{\text{train}})$ as proper priors, we can define the following *partial* Bayes factor

$$\text{BF}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}}) = \frac{Z^{(1)}_{\text{test}|\text{train}}}{Z^{(2)}_{\text{test}|\text{train}}} = \frac{\frac{Z_1}{Z^{(1)}_{\text{train}}}}{\frac{Z_2}{Z^{(2)}_{\text{train}}}} \tag{200}$$

$$= \frac{\frac{Z_1}{Z_2}}{\frac{Z^{(1)}_{\text{train}}}{Z^{(2)}_{\text{train}}}} = \frac{\text{BF}(\mathbf{y})}{\text{BF}(\mathbf{y}_{\text{train}})}. \quad \textit{(``Bayes law for Bayes Factors'').} \tag{201}$$

The last expression does not depend on c_1, c_2, since

$$\text{BF}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}}) = \frac{\text{BF}(\mathbf{y})}{\text{BF}(\mathbf{y}_{\text{train}})} = \frac{\frac{c_1 \tilde{Z}_1}{c_2 \tilde{Z}_2}}{\frac{c_1 \tilde{Z}^{(1)}_{\text{train}}}{c_2 \tilde{Z}^{(2)}_{\text{train}}}} = \frac{\tilde{Z}_1}{\tilde{Z}_2} \frac{\tilde{Z}^{(2)}_{\text{train}}}{\tilde{Z}^{(1)}_{\text{train}}}. \tag{202}$$

Therefore, one can approximate firstly $\text{BF}(\mathbf{y}_{\text{train}})$, secondly $\text{BF}(\mathbf{y})$ and then compare the model using the partial Bayes factor $\text{BF}(\mathbf{y}_{\text{test}}|\mathbf{y}_{\text{train}})$.

Remark. The trick here consists in computing *two normalizing constants* for each model, instead of only one. The first normalizing constant is used for building an auxiliary normalized prior, depending on $\mathbf{y}_{\text{train}}$.

A training dataset $\mathbf{y}_{\text{train}}$ is proper if $\int_{\mathcal{X}_m} \ell_m(\mathbf{y}_{\text{train}}|\mathbf{x}_m) h_m(\mathbf{x}_i) d\mathbf{x}_m < \infty$ for all models, and it is called *minimal* if is proper and no subset of $\mathbf{y}_{\text{train}}$ is proper. If we use actually proper prior densities, the minimal training dataset is the empty set and the fractional Bayes factor reduces to the classical Bayes factor. However, the main drawback of the partial Bayes factor approach is the dependence on the choice of $\mathbf{y}_{\text{train}}$ (which could affect the selection of the model). The authors suggest to find the *minimal* suitable training set $\mathbf{y}_{\text{train}}$, but this task is not straightforward. Two alternatives in the literature have been proposed, the fractional Bayes factors and the intrinsic Bayes factors.

Fractional Bayes Factors [69]. Instead of using a training data, it is possible to use power posteriors, i.e.,

$$\text{FBF}(\mathbf{y}) = \frac{\text{BF}(\mathbf{y})}{\text{BF}(\mathbf{y}|\beta)}, \tag{203}$$

where the denominator is

$$\text{BF}(\mathbf{y}|\beta) = \frac{\int_{\mathcal{X}_1} \ell_1(\mathbf{y}|\mathbf{x}_1)^\beta g_1(\mathbf{x}_1) d\mathbf{x}_1}{\int_{\mathcal{X}_2} \ell_2(\mathbf{y}|\mathbf{x}_2)^\beta g_2(\mathbf{x}_2) d\mathbf{x}_2} = \frac{c_1 \int_{\mathcal{X}_1} \ell_1(\mathbf{y}|\mathbf{x}_1)^\beta h_1(\mathbf{x}_1) d\mathbf{x}_1}{c_2 \int_{\mathcal{X}_2} \ell_2(\mathbf{y}|\mathbf{x}_2)^\beta h_2(\mathbf{x}_2) d\mathbf{x}_2}. \tag{204}$$

with $0 < \beta \leq 1$, and $\mathrm{BF}(\mathbf{y}|1) = \mathrm{BF}(\mathbf{y})$. Note that the value $\beta = 0$ is not admissible since $\int_{\mathcal{X}_m} h_m(\mathbf{x}_m) d\mathbf{x}_m = \infty$ for $m = 1, 2$. Again, since both $\mathrm{BF}(\mathbf{y})$ and $\mathrm{BF}(\mathbf{y}|\beta)$ depend on the ratio $\frac{c_1}{c_2}$, the fractional Bayes factor $\mathrm{FBF}(\mathbf{y})$ is independent on c_1 and c_2 by definition.

Intrinsic Bayes factors [3]. The partial Bayes factor (200) will depend on the choice of (minimal) training set $\mathbf{y}_{\text{train}}$. These authors solve the problem of choosing the training sample by averaging the partial Bayes factor over all possible minimal training sets. They suggest to use the arithmetic mean, leading to the *arithmetic intrinsic* Bayes factor, or the geometric mean, leading to the *geometric intrinsic* Bayes factor.

7 Theoretical and empirical comparisons

In this section, we illustrate the performance of different marginal likelihood estimators in different experiments. In Section 7.1, first we compare theoretically the variance of IS and RIS estimators in a one-dimensional example, which allows to discuss some important features required by the proposal and auxiliary densities (e.g., the conditions regarding the tails of these pdfs). In a second part of Section 7.1, we compare the Mean Square Error (MSE) and bias of IS an RIS estimators via numerical simulations. In Section 7.2 we test several estimators in a nonlinear regression problem with real data, where the likelihood function has highly non-elliptical contours.

7.1 Theoretical and empirical comparison of IS and RIS

In this example, our goal is to compare, theoretically and by numerical simulations, the standard IS and RIS schemes for estimating the normalizing constant of a Gaussian target $\pi(x) = \exp(-\frac{1}{2}x^2)$. We know the ground-truth $Z = \int_{-\infty}^{\infty} \pi(x) dx = \sqrt{2\pi}$, so $\bar{\pi}(x) = \frac{\pi(x)}{Z} = \mathcal{N}(x|0, 1)$. The standard IS estimator of Z with importance density $\bar{q}(x)$ and the RIS estimator with auxiliary density $f(x)$ are

$$\widehat{Z}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(x_i)}{\bar{q}(x_i)}, \quad x_i \sim \bar{q}(x), \quad \widehat{Z}_{\text{RIS}} = \frac{1}{\frac{1}{N} \sum_{i=1}^{N} \frac{f(x_i)}{\pi(x_i)}}, \quad x_i \sim \bar{\pi}(\mathbf{x}). \tag{205}$$

For a fair theoretical and empirical comparison, we consider

$$\bar{q}(x) = f(x) = \mathcal{N}(x|0, h^2) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{1}{2h^2} x^2\right). \tag{206}$$

where $h > 0$ is the standard deviation. We desire to study the performance of the two estimators as h varies.

7.1.1 Theoretical analysis

We compute the variances of the estimators $\widehat{Z}_{\text{IS}}$ and $\widehat{Z}_{\text{RIS}}$ as functions of h, starting from $\widehat{Z}_{\text{IS}}$. Note that by the i.i.d. assumption, we can write

$$\text{var}[\widehat{Z}_{\text{IS}}] = \frac{1}{N}\text{var}_{\bar{q}}\left[\frac{\pi(x)}{\bar{q}(x)}\right] = \frac{1}{N}\left\{\mathbb{E}_{\bar{q}}\left[\frac{\pi(x)}{\bar{q}(x)}\right]^2 - Z^2\right\}, \tag{207}$$

Substituting $\pi(x) = \exp(-\frac{1}{2}x^2)$ and $\bar{q}(x) = \frac{1}{\sqrt{2\pi h^2}}\exp(-\frac{1}{2h^2}x^2)$ we obtain

$$\begin{aligned} \mathbb{E}_{\bar{q}}\left[\frac{\pi(x)}{\bar{q}(x)}\right]^2 &= \int_{-\infty}^{\infty}\left(\frac{\pi(x)}{\bar{q}(x)}\right)^2 \bar{q}(x)dx && (208)\\ &= \int_{-\infty}^{\infty}\frac{\pi(x)^2}{\bar{q}(x)}dx && (209)\\ &= \sqrt{2\pi h^2}\int_{-\infty}^{\infty}\exp\left\{-\left(1-\frac{1}{2h^2}\right)x^2\right\}dx && (210)\\ &= 2\pi\frac{h}{\sqrt{2-\frac{1}{h^2}}}. && (211)\end{aligned}$$

Replacing the last expression above in Eq. (207), we obtain that the variance of $\widehat{Z}_{\text{IS}}$ is given by

$$\text{var}_{\bar{q}}\left[\widehat{Z}_{\text{IS}}\right] = \frac{2\pi}{N}\left\{\frac{h}{\sqrt{2-\frac{1}{h^2}}} - 1\right\}. \tag{212}$$

This variance reaches its minimum, var[$\widehat{Z}_{\text{IS}}$] $= 0$, at $h = 1$, i.e., when the proposal is exactly the posterior $\bar{q}(x) = \mathcal{N}(x|0,1) = \bar{\pi}(x)$, as expected. For $h < 1$, var[$\widehat{Z_{\text{IS}}}$] grows exponentially until reaching $h = \frac{1}{\sqrt{2}}$ where is infinite. For $0 < h < \frac{1}{\sqrt{2}}$, var[$\widehat{Z_{\text{IS}}}$] is not defined. Finally, var[$\widehat{Z_{\text{IS}}}$] grows linearly from $h = 1$ onwards, i.e., diverges to infinity as $h \to \infty$. Figure 3-(a) shows that behavior, with $N = 500$. Note that changing the value of N simply scales the whole curve. Clearly, this is perfectly in line the well-known theoretical requirement that the proposal pdf must have fatter tails than the posterior density in a IS scheme. Moreover, this confirms that the use of proposals with variance bigger than that of the target is generally not catastrophic. The opposite could yield catastrophic results. Recall also that $\mathbb{E}_{\bar{q}}[\widehat{Z}_{\text{IS}}] = Z$, i.e., the bias of $\widehat{Z}_{\text{IS}}$ is zero.

Regarding RIS, it is easier to compute the variance of $\widehat{r} = \frac{1}{\widehat{Z}_{\text{RIS}}}$, rather than $\widehat{Z}_{\text{RIS}}$ itself. Namely, we consider the estimator $\widehat{r} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(x_i)}{\pi(x_i)}$, with $x_i \sim \bar{\pi}(x)$. which is an unbiased estimator of $\frac{1}{Z}$. Since x_i's are i.i.d. from $\bar{\pi}(x)$, then we have

$$\begin{aligned} \text{var}_{\bar{\pi}}\left[\widehat{r}\right] = \text{var}_{\bar{\pi}}\left[\frac{1}{\widehat{Z}_{\text{RIS}}}\right] &= \frac{1}{N}\text{var}_{\bar{\pi}}\left[\frac{f(x)}{\pi(x)}\right] && (213)\\ &= \frac{1}{N}\left\{\mathbb{E}_{\bar{\pi}}\left[\frac{f(x)}{\pi(x)}\right]^2 - \frac{1}{Z^2}\right\}, && (214)\end{aligned}$$

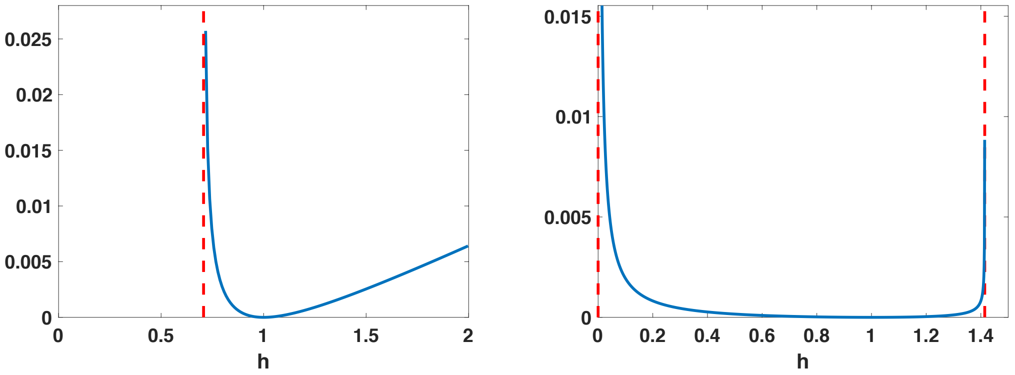
Substituting $\pi(x) = \exp\left(-\frac{1}{2}x^2\right)$ and $f(x) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{1}{2h^2}x^2\right)$ we obtain

$$\begin{aligned}
\mathbb{E}_{\bar{\pi}}\left[\frac{f(x)}{\pi(x)}\right]^2 &= \int_{-\infty}^{\infty} \left(\frac{f(x)}{\pi(x)}\right)^2 \frac{\pi(x)}{Z} dx && (215)\\
&= \frac{1}{Z}\int_{-\infty}^{\infty} \frac{f(x)^2}{\pi(x)} dx && (216)\\
&= \frac{1}{2\pi h^2\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{1}{h^2}-\frac{1}{2}\right)x^2\right\} dx && (217)\\
&= \frac{1}{2\pi}\frac{1}{h^2\sqrt{\frac{2}{h^2}-1}}. && (218)
\end{aligned}$$

Hence the variance of $\widehat{r}$ is given by

$$\text{var}_{\bar{\pi}}[\widehat{r}] = \text{var}_{\bar{\pi}}\left[\frac{1}{\widehat{Z}_{\text{RIS}}}\right] = \frac{1}{2\pi N}\left\{\frac{1}{h^2\sqrt{\frac{2}{h^2}-1}} - 1\right\}, \tag{219}$$

which reaches its minimum, $\text{var}_{\bar{\pi}}[\widehat{r}] = 0$, at $h = 1$ as expected. Note that $\text{var}[\widehat{r}]$ is defined when $0 < h < \sqrt{2}$ (there are two vertical asymptotes). Moreover, $\text{var}_{\bar{\pi}}[\widehat{r}]$ grows more quickly in $1 < h < \sqrt{2}$ than in $0 < h < 1$. In Figure 3-(b), we show $\text{var}_{\bar{\pi}}[1/\widehat{Z}_{\text{RIS}}]$ for $N = 500$. Again, the effect of the number of samples N is simply a scaling factor of the curve. Indeed, $\widehat{r} = \frac{1}{\widehat{Z}_{\text{RIS}}}$ has the same behavior as the IS estimator when the variance of the denominator (in this case $\bar{\pi}(x)$) is smaller than the numerator (in this case $f(x)$). Then, we see that $f(x)$ should have non-zero variance and less variance than $\bar{\pi}(x)$, in order to avoid infinite variance of the resulting estimator $\widehat{r} = \frac{1}{\widehat{Z}_{\text{RIS}}}$. We can observe two vertical asymptotes when we analyze $\text{var}_{\bar{\pi}}[1/\widehat{Z}_{\text{RIS}}]$. However, note that $\text{var}_{\bar{\pi}}[\widehat{Z}_{\text{RIS}}]$ has just one vertical asymptote at $h = 0$ as shown below.



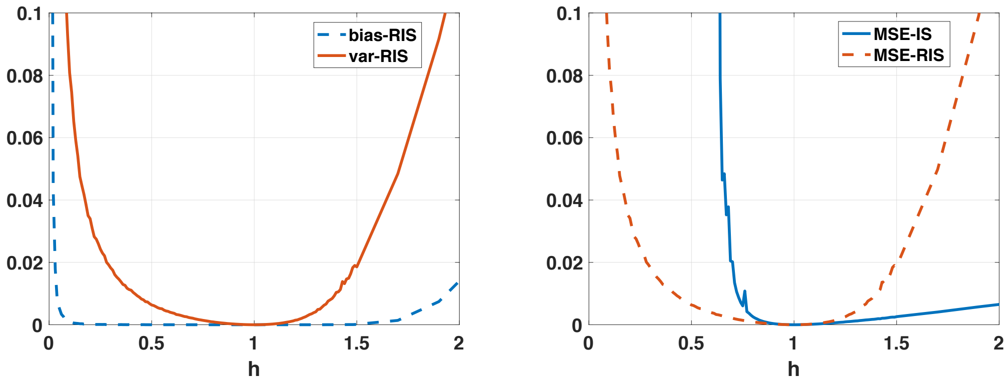
(a) Variance of $\widehat{Z}_{\text{IS}}$.

(b) Variance of $\frac{1}{\widehat{Z}_{\text{RIS}}}$.

Figure 3: The variances $\text{var}_{\bar{\pi}}[\widehat{Z}_{\text{IS}}]$ and $\text{var}_{\bar{\pi}}[1/\widehat{Z}_{\text{RIS}}]$ in Eqs. (212) and (219), respectively ($N = 500$).

7.1.2 Numerical analysis

Setting $N = 500$, we compute numerically the mean square error (MSE) of both $\widehat{Z}_{\text{IS}}$ and $\widehat{Z}_{\text{RIS}}$, and the variance and bias of both $\widehat{Z}_{\text{RIS}}$, averaging the results over 5000 independent runs. We show the results in Fig. 7.1.2. In Figure 7.1.2(a), we show bias and variance of the estimator $\widehat{Z}_{\text{RIS}}$. Note that its variance has only one asymptote at 0 instead of two asymptotes, unlike the variance of $\widehat{r} = 1/\widehat{Z}_{\text{RIS}}$. Also the bias diverges in 0. Observe also that the bias is negligible for $0.1 < h < 1.6$, with respect to the value of the variance. In Fig. 7.1.2-(b), we can see that the MSE of $\widehat{Z}_{\text{IS}}$ corresponds to its theoretical variance shown in Fig. 3, as we expect since $\widehat{Z}_{\text{IS}}$ has zero bias, hence $\text{MSE}(\widehat{Z}_{\text{IS}}) = \text{var}(\widehat{Z}_{\text{IS}})$. Although $\widehat{Z}_{\text{RIS}}$ is not unbiased, we see that its MSE, also shown in Fig. 7.1.2-(b), is virtually identical to its variance shown in Fig. 7.1.2-(a), where the bias seems to be negligible for the majority of values of h.



(a) Bias and variance of $\widehat{Z}_{\text{RIS}}$.

(b) MSE of $\widehat{Z}_{\text{IS}}$ and $\widehat{Z}_{\text{RIS}}$.

Figure 4: (a) Bias (dashed line) and variance (solid line) of $\widehat{Z}_{\text{RIS}}$ as a function of h ($N = 500$). (b) MSE of $\widehat{Z}_{\text{IS}}$ (solid line) and $\widehat{Z}_{\text{RIS}}$ (dashed line) as a function of h ($N = 500$).

7.2 Experiment with biochemical oxygen demand data

We consider a numerical experiment studied also in [19], that is a nonlinear regression problem modeling data on the biochemical oxygen demand (BOD) in terms of time instants. The outcome variable Y_i = BOD (mg/L) is modeled in terms of t_i = time (days) as

$$Y_i = x_1(1 - e^{-x_2 t_i}) + \epsilon_i, \quad i = 1, \dots, 6, \tag{220}$$

where the ϵ_i's are independent $\mathcal{N}(0, \sigma^2)$ errors, hence $Y_i \sim \mathcal{N}(x_1(1 - e^{-x_2 t_i}), \sigma^2)$. The data $\mathbf{y} \equiv \{y_i\}_{i=1}^6$, measured at locations $\{t_i\}_{i=1}^6$, are shown in Table 14 below.
The goal is to compute the normalizing constant of the posterior of $\mathbf{x} = (x_1, x_2)$ given the data $\mathbf{y}$. Following [19], we consider uniform priors for $x_1 \sim \mathcal{U}([0, 60])$, and $x_2 \sim \mathcal{U}([0, 6])$, i.e., $g_1(x_1) = \frac{1}{60}$ for $x_1 \in [0, 60]$, and $g_2(x_2) = \frac{1}{6}$, with $x_2 \in [0, 6]$. Moreover, we consider an improper prior for

Table 14: Data of the numerical experiment in Section 7.2.

t_i (days)	y_i (mg/L)
1	8.3
2	10.3
3	19.0
4	16.0
5	15.6
7	19.8

σ, $g_3(\sigma) \propto \frac{1}{\sigma}$. However, we will integrate out the variable σ. Indeed, the two-dimensional target $\pi(\mathbf{x}) = \pi(x_1, x_2)$ results after integrating out σ by marginalizing

$$\pi(x_1, x_2, \sigma) = \ell(\mathbf{y}|x_1, x_2, \sigma)g_1(x_1)g_2(x_2)g_3(\sigma),$$

w.r.t. σ, namely we obtain

$$\pi(\mathbf{x}) = \int \pi(x_1, x_2, \sigma)d\sigma = \ell\left(\mathbf{y}|x_1, x_2\right) g_1(x_1)g_2(x_2) \tag{221}$$

$$= \frac{1}{60}\frac{1}{6}\frac{1}{\pi^3}\frac{8}{\left\{\sum_{i=1}^{6}[y_i - x_1(1-\exp(-x_2 t_i))]^2\right\}^3}, \quad (x_1, x_2) \in [0, 60] \times [0, 6], \tag{222}$$

for which we want to compute its normalizing constant $Z = \int \pi(\mathbf{x})d\mathbf{x}$. The derivation is given in the Supplementary Material. The true value (ground-truth) is $\log Z = -16.208$, considering the data in Table 14. As in [19], we compare the relative mean absolute error

$$\frac{\mathbb{E}\left[|\widehat{Z} - Z|\right]}{Z},$$

obtained by different methods:

- the naive Monte Carlo estimator,
- a modified version of the Laplace method (more sophisticated) given in [19],
- a "crude" Laplace scheme (using sample mean and sample covariance considering MCMC samples from $\bar{\pi}$),
- the HM estimator of Eq. 46,
- the RIS estimator where $f(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the MCMC samples from $\bar{\pi}$ (it is denoted as RIS in Table 15),
- another RIS scheme where $f(\mathbf{x})$ is obtained by a KDE with $K = 4$ clusters and $h = 0$ (in a similar fashion of Eq. (152)),

- an IS estimator with a Gaussian proposal pdf $\bar{q}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where again $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the MCMC samples from the posterior,
- and, finally, a CLAIS scheme with $K = 2$, $h = 0$ i.e., as in Eq. (152).

Table 15: Relative error, and its corresponding standard error, in estimating the marginal likelihood by seven methods

	Naive	Laplace (soph)	Laplace	HM	RIS	RIS-kde	IS	CLAIS
RE	0.057	0.181	0.553	0.823	0.265	0.140	0.084	0.082
std err	0.001	0.013	0.003	0.018	0.006	0.004	0.015	0.014
comments	—	see [19]	—	—	—	$K = 4$	—	$K = 2$

To obtain the samples from the posterior, we run $T = 10000$ iterations of a Metropolis-Hastings algorithm, using the prior as an independent proposal pdf. Since CLAIS draws additional samples from $\bar{q}(\mathbf{x})$ in the lower layer, in order to provide a fair comparison, in CLAIS we consider $N = 1$, $T' = T/2 = 5000$ and 5000 additional samples in the lower layer. We averaged the relative error over 1000 independent runs. Our results are shown in Table 15.

In this examples, and with these priors, the results show that the best performing estimator in this case is the Naive Monte Carlo, since prior and likelihood has an ample overlapping region of probability mass. However, the naive Monte Carlo scheme is generally inefficient when there is a small overlap between likelihood and prior. Note also that IS and CLAIS provide good performance. The worst performance is provided by the HM estimator.

8 Final discussion

In this work, we have provided an exhaustive review of the techniques for marginal likelihood computation with the purpose of model selection and hypothesis testing. Methods for approximating ratios of normalizing constants have been also described. A careful use of the improper priors in the Bayesian setting has been discussed. Most of the presented techniques are based on the importance sampling (IS) approach, but also require the use of MCMC algorithms. Table 16 summarizes some methods for estimating Z, that can be employed if N samples from the posterior are available. This table is devoted to the interested readers which desire to obtain samples $\{\mathbf{x}_n\}_{n=1}^N$ by an MCMC method with invariant pdf $\bar{\pi}(\mathbf{x})$ (without either any tempering or sequence of densities) and, at the same time, also desire to approximate Z using $\{\mathbf{x}_n\}_{n=1}^N$. Clearly, this table provides only a subset of all the possible techniques. They can be considered the simplest schemes, in the sense that they do not use any tempering strategy or sequence of densities. We also recall that AIC and DIC are commonly used for model comparison, although they do not directly target the actual marginal likelihood.

Table 16: Schemes for estimating Z, after that N samples have been generated by an MCMC algorithm with invariant distribution $\bar{\pi}(\mathbf{x})$.

Method	**Section**	**Need of drawing additional samples**	**Comments**
Laplace	2	—	use MCMC for estimating $\widehat{\mathbf{x}}_{\text{MAP}}$
BIC	2	—	use MCMC for estimating $\widehat{\mathbf{x}}_{\text{MLE}}$
KDE	2	—	use MCMC for generating samples
Chib's method	2	✓	additional samples are required if the proposal is not independent
RIS	3	—	the HM estimator is a special case
MTM	4.2	—	provides two estimators of Z
[78]	4.3	—	related to LAIS
LAIS	4.3	✓	with $\bar{\pi}$ in the upper-layer
Below: for model selection but do not approximate the marginal likelihood			
AIC	2	—	use MCMC for estimating $\widehat{\mathbf{x}}_{\text{MLE}}$
DIC	2	—	use MCMC for estimating c_p and $\bar{\mathbf{x}}$

References

[1] D. Ardia, N. Baştürk, L. Hoogerheide, and H. K. Van Dijk. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics & Data Analysis*, 56(11):3398–3414, 2012.

[2] V. Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural computation*, 9(2):349–368, 1997.

[3] J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[4] C. S. Bos. A comparison of marginal likelihood computation methods. In *Compstat*, pages 111–116. Springer, 2002.

[5] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

[6] M. F. Bugallo, L. Martino, and J. Corander. Adaptive importance sampling in signal processing. *Digital Signal Processing*, 47:36–49, 2015.

[7] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

[8] B. P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484, 1995.

[9] M.-H. Chen, Q.-M. Shao, et al. On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 1997.

[10] M. H. Chen, Q. M. Shao, and J. G. Ibrahim. *Monte Carlo methods in Bayesian computation.* Springer, 2012.

[11] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

[12] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

[13] N. Chopin. A sequential particle filter for static models. *Biometrika*, 89:539–552, 2002.

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

[15] G. Claeskens and N. L. Hjort. The focused information criterion. *Journal of the American Statistical Association*, 98(464):900–916, 2003.

[16] P. Congdon. Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational statistics & data analysis*, 50(2):346–357, 2006.

[17] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, December 2012.

[18] P. Dellaportas, J. J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002.

[19] T. J. DiCiccio, R. E. Kass, A. Raftery, and L. Wasserman. Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915, 1997.

[20] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez. Particle filtering. *IEEE Signal Processing Magazine*, 20(5):19–38, September 2003.

[21] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. *technical report*, 2008.

[22] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Efficient multiple importance sampling estimators. *IEEE Signal Processing Letters*, 22(10):1757–1761, 2015.

[23] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Heretical multiple importance sampling. *IEEE Signal Processing Letters*, 23(10):1474–1478, 2016.

[24] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized Multiple Importance Sampling. *Statistical Science*, 34(1):129–155, 2019.

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

[26] N. Friel and J. Wyse. Estimating the evidence-a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

[27] A. E. Gelfand and D. K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

[28] A. Gelman and X. L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

[29] W. R. Gilks and C. Berzuini. Following a moving target-Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):127–146, 2001.

[30] W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Applied Statistics*, 44(4):455–472, 1995.

[31] W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice.* Chapman and Hall/CRC, 1995.

[32] W. R. Gilks and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, 41(2):337–348, 1992.

[33] S. J. Godsill. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of computational and graphical statistics*, 10(2):230–248, 2001.

[34] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[35] Q.F. Gronau, H. Singmann, and E.J. Wagenmakers. Bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors. *R package version 0.4-0, URL https://CRAN. R-project. org/package= bridgesampling*, 2017.

[36] P. Grunwald and T. Roos. Minimum Description Length Revisited. *arXiv:1908.08484*, pages 1–38, 2019.

[37] D. I. Hastie and P. J. Green. Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338, 2012.

[38] J. A. Hoeting, D. Madigan, A. E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

[39] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

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

[41] A. Kong. A note on importance sampling using standardized weights. *Technical Report 348, Department of Statistics, University of Chicago*, 1992.

[42] C. H. LaMont and P. A. Wiggins. Correspondence between thermodynamics and inference. *Physical Review E*, 99(5):052140, 2019.

[43] S. M. Lewis and A. E. Raftery. Estimating Bayes factors via posterior simulation with the LaplaceMetropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, 1997.

[44] F. Liang, C. Liu, and R. Caroll. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley Series in Computational Statistics, England, 2010.

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

[46] P.i Liu, A. S. Elshall, M. Ye, P. Beerli, X. Zeng, D. Lu, and Y. Tao. Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. *Water Resources Research*, 52(2):734–758, 2016.

[47] J. M. Marin and C. P. Robert. Importance sampling methods for Bayesian discrimination between embedded models. *arXiv preprint arXiv:0910.2325*, 2009.

[48] A. D. Martin, K. M. Quinn, and J. H. Park. *MCMCpack: Markov chain Monte Carlo in R*. Foundation for Open Access Statistics, 2011.

[49] L. Martino. A review of multiple try MCMC algorithms for signal processing. *Digital Signal Processing*, 75:134 – 152, 2018.

[50] L. Martino, R. Casarin, F. Leisen, and D. Luengo. Adaptive independent sticky MCMC algorithms. *EURASIP Journal on Advances in Signal Processing (to paper)*, 2017.

[51] L. Martino and V. Elvira. Metropolis sampling. *Wiley StatsRef: Statistics Reference Online*, pages 1–18, 2017.

[52] L. Martino and V. Elvira. Compressed Monte Carlo for distributed Bayesian inference. *viXra:1811.0505*, 2018.

[53] L. Martino, V. Elvira, and G. Camps-Valls. Group importance sampling for particle filtering and MCMC. *Digital Signal Processing*, 82:133 – 151, 2018.

[54] L. Martino, V. Elvira, and F. Louzada. Weighting a resampled particle in Sequential Monte Carlo. *IEEE Statistical Signal Processing Workshop, (SSP)*, 122:1–5, 2016.

[55] L. Martino, V. Elvira, and M. F. Louzada. Effective Sample Size for importance sampling based on the discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[56] L. Martino, V. Elvira, and D. Luengo. Anti-tempered layered adaptive importance sampling. *International Conference on Digital Signal Processing (DSP)*, 2017.

[57] L. Martino, V. Elvira, D. Luengo, and J. Corander. Layered adaptive importance sampling. *Statistics and Computing*, 27(3):599–623, 2017.

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

[59] L. Martino, V. P. Del Olmo, and J. Read. A multi-point Metropolis scheme with generic weight functions. *Statistics & Probability Letters*, 82(7):1445–1453, 2012.

[60] L. Martino and J. Read. On the flexibility of the design of multiple try Metropolis schemes. *Computational Statistics*, 28(6):2797–2823, 2013.

[61] L. Martino, J. Read, V. Elvira, and F. Louzada. Cooperative parallel particle filters for on-line model selection and applications to urban mobility. *Digital Signal Processing*, 60:172–185, 2017.

[62] L. Martino, J. Read, and D. Luengo. Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling. *IEEE Transactions on Signal Processing*, 63(12):3123–3138, June 2015.

[63] X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

[64] A. Mira and G. Nicholls. Bridge estimation of the probability density at a point. Technical report, Department of Mathematics, The University of Auckland, New Zealand, 2003.

[65] A. Mira and G. Nicholls. Bridge estimation of the probability density at a point. Technical report, Department of Mathematics, The University of Auckland, New Zealand, 2003.

[66] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[67] C. A. Naesseth, F. Lindsten, and T. B. Schon. Nested Sequential Monte Carlo methods. *Proceedings of theInternational Conference on Machine Learning*, 37:1–10, 2015.

[68] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[69] A. O'Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–118, 1995.

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

[71] C. M. Pooley and G. Marion. Bayesian model evidence as a practical alternative to deviance information criterion. *Royal Society Open Science*, 5(3):1–16, 2018.

[72] J. R Oaks, K. A. Cobb, V. N Minin, and A. D. Leaché. Marginal likelihoods in phylogenetics: a review of methods and applications. *Systematic biology*, 68(5):681–697, 2019.

[73] C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. *Advances in neural information processing systems*, pages 505–512, 2003.

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

[75] C. P. Robert and D. Wraith. Computational methods for Bayesian model choice. *AIP conference proceedings*, 1193(1):251–262, 2009.

[76] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81, 1986.

[77] A.i Schöniger, T. Wöhling, L. Samaniego, and W. Nowak. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water resources research*, 50(12):9484–9513, 2014.

[78] I. Schuster and I. Klebanov. Markov Chain Importance Sampling-a highly efficient estimator for MCMC. *arXiv preprint arXiv:1805.07179*, 2018.

[79] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

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

[81] D. J. Spiegelhalter and A. F.M. Smith. Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):377–387, 1982.

[82] D.J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B*, 64:583–616, 2002.

[83] D.J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. The deviance information criterion: 12 years on. *J. R. Stat. Soc. B*, 76:485–493, 2014.

[84] Statisticat and LLC. *LaplacesDemon: Complete Environment for Bayesian Inference*, 2018. R package version 16.1.1.

[85] G.l Stoltz and M. Rousset. Free energy computations: A mathematical perspective. *World Scientific*, 2010.

[86] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[87] I. Urteaga, M. F. Bugallo, and P. M. Djurić. Sequential Monte Carlo methods under model uncertainty. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5, 2016.

[88] V. Vyshemirsky and M. A. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2007.

[89] M. D. Weinberg et al. Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution. *Bayesian Analysis*, 7(3):737–770, 2012.

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