Statistics and Computing

Numerical daemons in Monte Carlo estimators of Bayesian model evidence

--Manuscript Draft--

Manuscript Number:	
Full Title:	Numerical daemons in Monte Carlo estimators of Bayesian model evidence
Article Type:	Manuscript
Keywords:	Bayesian model evidence, rounding error, path sampling, importance sampling, Monte Carlo simulation, Markov chain Monte Carlo
Corresponding Author:	Ming Ye Florida State University Tallahassee, FL UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Florida State University
Corresponding Author's Secondary Institution:	
First Author:	Ahmed S Elshall
First Author Secondary Information:	
Order of Authors:	Ahmed S Elshall
	Ming Ye
Order of Authors Secondary Information:	
Funding Information:	directorate for geosciences (EAR-1552329) — Dr. Ming Ye
Abstract:	Estimating Bayesian model evidence (BME) is of central importance in Bayesian multimodel analyses such as Bayesian model averaging and model selection. Due to their increased accuracy, the Monte-Carlo-based numerical estimators of BME are gaining attention over semi-analytical solutions (e.g., Laplace approximation). While the numerical estimators hold no theoretical assumptions about underlying distributions of BME, numerical implementation of the estimators can introduce several error sources. This work sheds light on two rounding sources that are rarely discussed in detail in literature. One is acceptance ratio rounding error that occurs during Markov chain Monte Carlo (MCMC) sampling, and the other is likelihood rounding error that occurs after the sample. These two numerical errors arise from arithmetic precision of floating point numbers. This manuscript illustrates the conditions that invoke these numerical errors, and demonstrates their implications on BME estimation. We show that the likelihood rounding error can result in severe underestimation of BME, while the acceptance ratio rounding error can result in overestimation of BME. We illustrate these two error sources using two popular estimators, namely, arithmetic mean (AM) and harmonic mean (HM), and two state-of-the-art estimators, namely, thermodynamic integration (TI) and steppingstone sampling (SS). We conceptually and numerically show how HM and SS are less vulnerable to these two error sources. Similarly, we illustrate how AM and TI are vulnerable to these numerical errors, resulting in underestimation and overestimation, respectively. We refer to these numerical errors as numerical daemons, as a daemon in computational language refers to a dormant process that is invoked when some conditions occur. By learning about these numerical daemons and the conditions that invoke them, users of BME numerical estimators can be aware of their adverse effects, and more advanced algorithms can be developed to avoid them.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Manuscript Click here to access/download;Manuscript;Manuscript_Final.pdf

Click here to view linked References

Numerical daemons in Monte Carlo estimators of Bayesian model evidence

Ahmed S. Elshall[1] and Ming Ye[1*]

[1]Department of Earth, Ocean and Atmospheric Science, Florida State University, Tallahassee, FL, USA

*Corresponding Author, email: mye@fsu.edu

Submit for Publication in Statistics and Computing

July 2021

Abstract

Estimating Bayesian model evidence (BME) is of central importance in Bayesian multimodel analyses such as Bayesian model averaging and model selection. Due to their increased accuracy, the Monte-Carlo-based numerical estimators of BME are gaining attention over semi-analytical solutions (e.g., Laplace approximation). While the numerical estimators hold no theoretical assumptions about underlying distributions of BME, numerical implementation of the estimators can introduce several error sources. This work sheds light on two rounding sources that are rarely discussed in detail in literature. One is acceptance ratio rounding error that occurs during Markov chain Monte Carlo (MCMC) sampling, and the other is likelihood rounding error that occurs after the sample. These two numerical errors arise from arithmetic precision of floating point numbers. This manuscript illustrates the conditions that invoke these numerical errors, and demonstrates their implications on BME estimation. We show that the likelihood rounding error can result in severe underestimation of BME, while the acceptance ratio rounding error can result in overestimation of BME. We illustrate these two error sources using two popular estimators, namely, arithmetic mean (AM) and harmonic mean (HM), and two state-of-the-art estimators, namely, thermodynamic integration (TI) and steppingstone sampling (SS). We conceptually and numerically show how HM and SS are less vulnerable to these two error sources. Similarly, we illustrate how AM and TI are vulnerable to these numerical errors, resulting in underestimation and overestimation, respectively. We refer to these numerical errors as numerical daemons, as a *daemon* in computational language refers to a dormant process that is invoked when some conditions occur. By learning about these numerical daemons and the conditions that invoke them, users of BME numerical estimators can be aware of their adverse effects, and more advanced algorithms can be developed to avoid them.

1. Introduction

Bayesian model evidence (BME) that is also known as normalizing constant, integrated likelihood, marginal likelihood, marginal evidence, model evidence, and evidence is defined as

$$p\left(\mathbf{D}\middle|M\right)=\int p\left(\mathbf{D}\middle|\boldsymbol{\theta},M\right)p\left(\boldsymbol{\theta}\middle|M\right)d\boldsymbol{\theta}, \quad (1)$$

where $p\left(\mathbf{D}\middle|\boldsymbol{\theta},M\right)$ is the joint likelihood of model M and its parameter $\boldsymbol{\theta}$ in reproducing the observation data $\mathbf{D}$, and $p\left(\boldsymbol{\theta}\middle|M\right)$ is the prior parameter probability density. The joint likelihood $p\left(\mathbf{D}\middle|\boldsymbol{\theta},M\right)$ states how likely the model would reproduce the observations for a realization of the model parameters. By marginalizing out the model parameters, BME states how likely the model would reproduce the observations given all the parameter values that the model can assume. Therefore, BME plays a critical role in Bayesian model averaging (Lu et al. 2015), model selection (Marshall et al. 2005), assessment of conceptual model and scenario uncertainty (Rojas et al. 2010), evaluating machine learning models (Chitsazan et al. 2015), evaluating ensemble prediction with pseudo Bayes factor (Elshall et al. 2018), optimum observation network design (Pham and Tsai 2016), data worth analysis (Wöhling et al. 2015), model complexity analysis (Schöniger et al. 2015), and many other applications. Given its vital importance, estimating BME is a common task in Bayesian computation.

Estimating BME is often challenging because it requires evaluating the multiple integral (Eq.1) over multi-dimensional model parameter space. This is often carried out using semi-analytical solutions (e.g., Laplace approximation, BIC, KIC, etc.) that involves theoretical simplifications, or using numerical methods that is based Monte Carlo (MC) methods that involve numerical approximations. As the assumptions of semi-analytical solutions may not be satisfied in practice, MC-based numerical methods can be a robust alternative (Friel and Wyse 2012; Liu et

al. 2016; Schöniger et al. 2014). Accordingly, there has been a trend in natural and social sciences to move toward numerical methods. However, a theoretically unbiased MC methods could have large bias in practice due to numerical errors. This manuscript focuses on one aspect of bias, which arises from rounding error, and we referred to as *numerical daemons.* The term "numerical daemon" was inspired by other numerical daemons papers (Clark and Kavetski 2010; Kavetski and Clark 2010; La Follette et al. 2021) in which numerical errors can heavily impact the solution. We show in detail that MC methods for estimating BME can be severally hampered by rounding error arising from arithmetic precision of floating point numbers. This precision is limited by the number digits that can be stored by the computer memory for each floating-point number. A value that is too large will overflow to *inf*, and a value that is too small will underflow to *zero*. This underflow and overflow due rounding error can have a large impact on BME estimation.

The objective of this study is to investigate the impact of rounding error on four MC estimators of BME. The first two estimators have been widely used in literature, and they approximate BME by calculating the arithmetic mean (AM, McCulloch and Rossi 1991) and harmonic mean (HM, Newton and Raftery 1994) of the joint likelihood, $p\left(\mathbf{D}\middle|\boldsymbol{\theta},M\right)$. The AM estimator uses parameter samples generated from prior parameter distribution $p\left(\boldsymbol{\theta}\middle|M\right)$. Although it is computationally expensive to sample the full prior distribution, the AM estimator is often used as a reference solution in benchmarking studies (e.g., Liu et al. 2016; Schöniger et al. 2014) because it is generally assumed that AM will approximate the true BME by using a large number of parameter samples. The HM estimator uses parameter samples generated from posterior parameter samples. As it is always necessary to sample the posterior, this makes the calculation of HM estimator computationally convenient at no extra computational cost (e.g., Zhang et al. 2014). However, it has been well known that the HM estimator is biased in practice (Calderhead and

Girolami 2009; Friel and Wyse 2012; Lartillot and Philippe 2006; Lefebvre et al. 2010; Liu et al. 2016; Xie et al. 2011). The other two MC estimators are the path sampling method of thermodynamic integration (TI, Friel and Pettitt 2008; Lartillot and Philippe 2006), and the importance sampling method of steppingstone sampling (SS, Xie et al. 2011). The key idea of the two methods is to use intermediate distributions along a sampling path that bridges from the prior to the posterior. Both TI and SS methods are the state-of-the-art methods for BME estimation in natural and social sciences. However, to appreciate the design of a good numerical estimator, a fundamental understanding of the inherent numerical errors in the MC estimators is necessary.

There are at least five sources of numerical error that can lead to biased BME estimation as summarized in Table 1. We briefly discuss these sources of error here, and explain them in detail through the rest of the manuscript. One error common to all MC methods is *MC stochastic error*. It decreases when sample size increase, as the law of large numbers ensures that MC estimate converges to the correct value as the number of samples increases. For the TI and SS estimators that utilize a sampling path of intermediate distributions bridging from the prior to the posterior, there exist a *discretization error,* which can be reduced by decreasing the Kullback-Liebler distance between successive intermediate distributions (Calderhead and Girolami 2009; Friel et al. 2014; Lefebvre et al. 2010). Generally speaking, for the path sampling based methods such as TI and SS, MC stochastic error and discretization error are the only two error sources that are discussed in literature (Gelman and Meng 1998; Lartillot and Philippe 2006; Xie et al. 2011).

Table 1. Sources of errors of Bayesian model evidence (BME) for the numerical estimators of arithmetic mean (AM), harmonic mean (HM), thermodynamic integration (TI) and steppingstone sampling (SS). The marks 'Yes' and 'No' represent that estimator is subject and not subject to this error source, respectively.

BME Error Source	Numerical Estimator			
	AM	HM	TI	SS
Stochastic sampling error	Yes	Yes	Yes	Yes
Discretization error	No	No	Yes	Yes
MCMC sampling error	No	Yes	Yes	Yes
Likelihood rounding error	Yes	Yes (unlikely)	No	Yes (unlikely)
Acceptance ratio rounding error	No	Yes (unlikely)	Yes	Yes

A third source of error arises when the Markov chain Monte Carlo (MCMC) algorithm cannot accurately approximate the target stationary distribution due to sampling difficulties, resulting in *MCMC sampling error*. For most practical problems all the intermediate distributions from prior to the posterior (including the posterior) are unknown and cannot be sampled directly using MC methods but by MCMC methods. Increased level of sampling difficulties (e.g., parameter correlation, multimodality, and high dimensionality) affects the MCMC sampling accuracy as illustrated in many studies (e.g., Laloy and Vrugt 2012). Elshall and Ye (2019) show that MCMC sampling errors can propagate in the intermediate distributions, which join the prior to posterior distributions. Elshall and Ye (2019) illustrate that MCMC sampling error can have higher impact on SS due to the effect of multiplication of intermediate distribution ratios than TI because the cumulative addition of the log-likelihood averages of intermediate distributions makes MCMC sampling errors less pronounced. Elshall and Ye (2019) propose the variant Multiple One Steppingstone Sampling (MOSS) that is less susceptible to MCMC sampling error, yet this variant cannot handle high-dimensional problems and peaked likelihood functions. In general, the MCMC sampling error can be reduced by developing more robust sampling algorithms and BME estimators. MCMC sampling error is beyond the scope of this work.

There are two additional sources of error that arise from rounding error, which has been realized by researchers but has not been discussed in detail, and they are the focus of this manuscript. We show that when using MC estimators to estimate BME, rounding error may introduce two numerical artifices that are the likelihood rounding error and the acceptance ratio rounding error. The *likelihood rounding error* occurs when we can accurately sample the target parameter distribution using MC simulation, but the correct likelihoods cannot be evaluated due to rounding error. On the other hand, the *acceptance ratio rounding error* occurs when MCMC simulation cannot accurately sample the target parameter values due to rounding error the acceptance ratio of the MCMC sampling. This work shows how likelihood rounding error is mainly associated with simple MC sampling (e.g., AM estimator), and acceptance ratio rounding error is mainly associated with MCMC sampling (e.g., HM, TI and SS estimators). Several studies (e.g., Baele et al. 2013; Fan et al. 2011; Lartillot and Philippe 2006) alluded to these numerical challenges when using MC estimators, yet without providing details. By studying the degree of vulnerability of AM, HM, TI, and SS estimators to rounding error, this study reveals two important facts that are often overlooked. First, we show that the likelihood rounding error causes AM to never converge to the true BME solution regardless of the computational effort of the simple MC sampling. In addition, we show that the acceptance ratio rounding error impacts the MCMC sampling. When using a likelihood function in MCMC simulations, this can result in trimming low likelihood regions and thus overestimation of BME. When using a log-likelihood function in MCMC simulations, which is generally considered to be more robust than the likelihood function, the rounding error may result in sampling the wrong stationary distributions. Although these might appear as tedious technical details, they provide valuable insights on interpreting the BME results of different MC estimators, especially when the results are contradicting.

The rest of the manuscript is organized as follows. In Section 2, we explain how rounding error can affect BME estimation, and give a theoretical overview about how the four MC estimators can deal with likelihood rounding error and the acceptance ratio rounding error. Section 3 illustrates the likelihood rounding error by using a high dimensional Gaussian problem with peaked likelihood function. This example explains why HM, TI and SS can handle likelihood rounding error, while AM cannot. In Section 4, we demonstrate that acceptance ratio rounding error occurs for even a simple two-dimensional linear problem with a relatively flat likelihood function. We use this example to show the impacts of acceptance ratio rounding error on MCMC sampling for both likelihood and log-likelihood functions, and explain why SS can potentially overcome this pitfall better than TI. Finally, we summarize our main findings and provide a fact-check list for a more transparent process of evaluating the results of BME estimation obtained using MC estimators.

2. Methodology

2.1 A path linking the prior and the posterior

One of the main ideas of path sampling, which is shared among TI and SS methods, is to define a path that links any two distributions. We will show later that AM and SS are two special cases. We can define a geometric path that links the prior to the unnormalized posterior. Define an unnormalized power posterior density as

$$q_{\beta_k}(\boldsymbol{\theta}) = p(\mathbf{D}|\boldsymbol{\theta}, M)^{\beta_k} p(\boldsymbol{\theta}|M), \quad (1)$$

where $p\left(\mathbf{D}\middle|\boldsymbol{\theta}, M\right)$ is likelihood of the parameter set $\boldsymbol{\theta}$ under model M to reproduce the observation data $\mathbf{D}$, and $p\left(\boldsymbol{\theta}\middle|M\right)$ is the prior densities of parameter $\boldsymbol{\theta}$ under model M. The power coefficient $\beta \in [0,1]$ is a scalar parameter for discretizing a continuous and differentiable path

linking two unnormalized power posterior densities. This path can be discretized by discrete power coefficients $0 = \beta_0 < \beta_1 < ... < \beta_{k-1} < \beta_k < ... < \beta_K = 1$. For $\beta_0 = 0$, Eq. 1 reduces to the prior $q_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid M)$, which is generally sampled directly using MC simulation. At $\beta_K = 1$, we have the unnormalized posterior $q_1(\boldsymbol{\theta}) = p(D \mid \boldsymbol{\theta}, M) p(\boldsymbol{\theta} \mid M)$, which is generally approximated and sampled using MCMC. At any other β values we obtain intermediate distributions that are less dispersed than the prior but more dispersed than the posterior (the readers may refer to Figure 3 for illustration purpose), and they are generally approximated and sampled using MCMC.

The unnormalized power posterior density $q_{\beta_k}(\boldsymbol{\theta})$ in Eq. 1 has the normalizing constant Z_{β_k} to yield the normalized power posterior density:

$$p_{\beta_k}(\boldsymbol{\theta} \mid \mathbf{D}, M) = \frac{q_{\beta_k}(\boldsymbol{\theta})}{Z_{\beta_k}} \tag{2a}$$

such that

$$Z_{\beta_k} = \int q_{\beta_k}(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2b}$$

From Eq. 2 we note two points. First, for $\beta = 1$ Eq. 2 reduces to Bayes' rule:

$$p_1\left(\boldsymbol{\theta} \middle| \mathbf{D}, M\right) = \frac{p\left(\mathbf{D} \middle| \boldsymbol{\theta}, M\right) p\left(\boldsymbol{\theta} \middle| M\right)}{p\left(\mathbf{D} \middle| M\right)} \tag{3}$$

and we have the BME as

$$p\left(\mathbf{D} \middle| M\right) = \int p\left(\mathbf{D} \middle| \boldsymbol{\theta}, M\right) p\left(\boldsymbol{\theta} \middle| M\right) d\boldsymbol{\theta} \tag{4}$$

such that $p_1\left(\boldsymbol{\theta} \middle| \mathbf{D}, M\right)$ is the posterior density. As Eq. 4 states how likely the model would reproduce the observation data given all the parameter values that the model can take, we can see that BME estimation is challenging because it involves evaluating a multidimensional integral of a usually highly nonlinear function that integrates over the whole parameter space. From Eq. 2 we

also note that at $\beta = 0$ the integral of the prior for $\boldsymbol{\theta}$ is the normalizing constant $Z_0 = 1$. At $\beta = 1$ the normalizing constant $Z_1 = p(\mathbf{D} \mid M)$ is integral of the likelihood for $\boldsymbol{\theta}$. Accordingly, the BME estimation amounts to estimating this ratio of normalizing constants:

$$p\left(\mathbf{D}\middle|M\right) = \frac{Z_1}{Z_0} \tag{5}$$

To compute this ratio TI and SS uses the path of distributions $q_\beta(\boldsymbol{\theta})_{0\le\beta\le1}$ defined by Eq. 1, AM uses the prior distribution at $\beta_0 = 0$, and HM uses the posterior distribution at $\beta_K = 1$. In the following discussion, we show the impact of rounding error on sampling these distributions, and how these four numerical estimators can handle rounding error in different ways.

2.2 Rounding error impact on likelihood and sampling

Evaluating the likelihood and MCMC sampling can be hampered by rounding error. Here we provide a brief account of rounding error and for detailed discussion the reader can refer to other sources (Heath 1997; Recktenwald 2000). Rounding error is the difference between result produced using exact arithmetic and result produced using limited precision arithmetic arising from rounding. In scientific computing terminology, rounding refers to the machine inexact representation of a real number as a floating-point number, and accordingly the real number is approximated by a nearby floating-point number. We limit our discussion here to the IEEE 754 standard double-precision binary floating-point arithmetic, but the discussion is valid in general. With the only notable exception of some mainframes and handheld calculators, IEEE 754 standard is adopted by most floating-point processors such as virtually all personal computers and workstations. For a processor that supports the IEEE 754 standard, the largest and smallest number that can be represent with double precision are negative real maximum $-R_{\max} \approx -1.7976 \times 10^{308}$ to negative real minimum $-R_{\min} \approx -2.225 \times 10^{-308}$, and real minimum $R_{\min} \approx 2.225 \times 10^{-308}$ to real

maximum $R_{\max} \approx 1.7976 \times 10^{308}$ with a floating-point relative accuracy $\varepsilon_{mach} = 2.2204 \times 10^{-16}$. The floating-point relative accuracy ε_{mach} (also known as machine precision and unit round-off) is the distance between a floating point and its nearest neighbor. Floating-point numbers larger than real maximum $R_{\max}$ or smaller than negative real maximum $-R_{\max}$ are assigned the values of positive and negative infinity, respectively. This is not the case for negative and positive real minimum $\pm R_{\min}$ as the gap between the smallest representable nonzero number and zero can be filled by additional floating-point numbers. These are referred to as subnormal or denormalized floating-point numbers. Giving the rule of rounding to nearest floating-point, the smallest nonzero denormalized number that can be presented is given by $\varepsilon_{dn} = \lim_{x \to 2}(R_{\min} \varepsilon_{mach} / x) \approx \pm 4.95 \times 10^{-324}$. Any positive or negative number smaller or larger than ε_{dn}, respectively, is assigned the value zero and stored in the computer memory as zero. This is referred to as *underflow* (also known as *floating point underflow* or *arithmetic underflow*). For the case of prior sampling using MC simulation, the rounding error can appear in the postprocessing of the likelihoods that is the likelihood rounding error. For any other distribution that requires MCMC sampling, the rounding error could occur during the sampling process, which is the acceptance ratio rounding error as explained in this section.

Given this limited precision arithmetic, all likelihood values smaller than ε_{dn} are assigned the value zero, and thus will be rejected by the Metropolis acceptance ratio that is given by

$$\alpha = \min\left(1, [\alpha_{power-posterior} \alpha_{prior}]\right) \tag{6a}$$

and the power posterior ratio is given by

$$\alpha_{power-posterior} = \left(\alpha_{posterior}\right)^{\beta_k} \tag{6b}$$

with

$$\alpha_{posterior} = \frac{L(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M)}{L(\mathbf{D} \mid \boldsymbol{\theta}_{old}, M)} \tag{6c}$$

such that the prior probability ratio $\alpha_{prior} = p(\boldsymbol{\theta}_{new} \mid M) / p(\boldsymbol{\theta}_{old} \mid M)$ is the ratio of the prior probability $p(\boldsymbol{\theta}_{new} \mid M)$ of the newly proposed sample $\boldsymbol{\theta}_{new}$, and the prior probability $p(\boldsymbol{\theta}_{old} \mid M)$ of the previously accepted sample $\boldsymbol{\theta}_{old}$; the posterior probability ratio $\alpha_{posterior}$ is the likelihood ratio of samples $\boldsymbol{\theta}_{new}$ and $\boldsymbol{\theta}_{old}$; and β_k is the power posterior coefficient as previous defined in Eq. 1 such that at $\beta_0 = 0$ we have the power posterior ratio $\alpha_{power-posterior} = 1$ for the prior distribution where all samples are accepted, and for $\beta_k > 0$ we obtain less dispersed distributions. In other words, from a thermodynamic perspective, the inverse of β_k is analogous to temperature such with higher temperature $\beta_k < 1$ we have more "looser" Markov chain that will visit and accept less likely parameter samples. Incrementally moving from $\beta_0 = 0$ to $\beta_K = 1$ is analogous to cooling down (i.e., annealing) such that we "tighten" the Markov chain and "narrow down" the likelihood surface. Given posterior probability ratio $\alpha_{posterior}$ as defined in Eq. 6, for a "looser" Markov chain we can obtain samples with likelihood values $L(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M)$ less than ε_{dn}. Their likelihoods will be assigned a values of zero and accordingly are continuously rejected. Thus, all probability spaces with $L(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M) < \varepsilon_{dn}$ cannot be explored.

One might argue that this discussion is irrelevant as log-likelihood is often favorably used instead for its numerical stability (Gelman and Meng 1998; Vrugt 2016). Actually, working with log-likelihood will lead to sampling with inaccurate Metropolis acceptance ratio

$$\alpha_{posterior} = \exp\left[\log(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M) - \log(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M)\right] \tag{7}$$

due to two limiting factors, which are both rounding error artifacts. The first arithmetic precision artifact is that the Metropolis acceptance ratio of log-likelihood is impaired by the smallest nonzero denormalized floating-point number $\log(\varepsilon_{dn}) = -745.1332$. From Eq. 7 it clear that when $\left[\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new}, M) - \log(\mathbf{D}\,|\,\boldsymbol{\theta}_{old}, M)\right] < \log(\varepsilon_{dn})$, taking the exponential yields $\alpha_{posterior} = 0$. Thus, all samples with $\alpha_{posterior} = 0$ will be rejected, which is similar to working with the likelihood function. The second arithmetic precision artifact is that the Metropolis acceptance ratio of log-likelihood is invalidly widened due to the numerical threshold imposed by the log real maximum $\log(R_{\max}) \approx 709.7827$. Given Eq. 7, when $\left[\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new}, M) - \log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new}, M)\right] > \log(R_{\max})$ taking exponential yields $\alpha_{posterior} = Inf$. This condition of producing a floating point number large than what a computer can store is referred to as *overflow* (also known as *floating point overflow* or *arithmetic overflow*). This indicates that working with log-likelihood can lead to accepting samples that should be rejected. These are samples that have high likelihood and low prior probability. Even though prior probability ratio α_{prior} could be extremely small stipulating that this sample should be rejected, substituting $\alpha_{posterior} = Inf$ in Eq. 6b yields $\alpha = Inf$. Thus, these samples are accepted at higher ratios than their true ratios. While working with likelihood function the Metropolis acceptance ratio is subject to only one threshold $\varepsilon_{dn} < \alpha_{posterior}$, when working with the log-likelihood function these two artifacts impose two thresholds $\varepsilon_{dn} < \alpha_{posterior} < R_{\max}$ on the Metropolis acceptance ratio. More importantly the second arithmetic precision artifact $\alpha_{posterior} < R_{\max}$ is much more critical because the first artifact $\varepsilon_{dn} < \alpha_{posterior}$ results in merely trimming part of prior parameter space, while the second artifact corrupts the sampling ratios.

2.3 Rounding error handling by BME estimators

The derivation of different BME estimators is beyond the scope of this work and the reader is referred other sources(Elshall and Ye 2019; Fan et al. 2011; Gelman and Meng 1998; Lartillot and Philippe 2006; Liu et al. 2016). Here we only focus on the basic concepts of importance sampling and path sampling techniques to provide a theoretical understanding of how different mathematical formulations of different estimators can handle rounding error. We start by importance sampling, which is the general technique used by AM, HM and SS to estimate the BME. Importance sampling evaluates an unknown distribution by using samples from an importance distribution in accordance to the importance weight. Using importance sampling technique, the BME can be estimated as (Xie et al. 2011):

$$p(\mathbf{D}\,|\,M)=\frac{E_g\left[p(\mathbf{D}\,|\,\boldsymbol{\theta},M)w(\boldsymbol{\theta}\,|\,M)\right]}{E_g\left[w(\boldsymbol{\theta}\,|\,M)\right]}=\frac{\frac{1}{n}\sum_{i=n}^{n}p(\mathbf{D}\,|\,\boldsymbol{\theta}_i,M)w_i(\boldsymbol{\theta}_i\,|\,M)}{\frac{1}{n}\sum_{i=n}^{n}w_i(\boldsymbol{\theta}_i\,|\,M)} \tag{8a}$$

using the importance distribution $g(\boldsymbol{\theta})$:

$$w(\boldsymbol{\theta}_i\,|\,M)=\frac{p(\boldsymbol{\theta}_i\,|\,M)}{g(\boldsymbol{\theta}_i)} \tag{8b}$$

where $p(\mathbf{D}\,|\,\boldsymbol{\theta}_i,M)$, $p(\boldsymbol{\theta}_i\,|\,M)$, $w(\boldsymbol{\theta}_i\,|\,M)$, and $g(\boldsymbol{\theta}_i)$ are the likelihood, prior density, importance weight, and the importance density, respectively, computed at parameter sample $\boldsymbol{\theta}_i$. By choosing different importance distributions in Eq. 8, we obtain different BME estimators. We consider how different approaches for solving Eq. 8, which are AH, HM and SS estimators, can handle numerical artifacts arising from rounding errors.

The arithmetic mean estimator is obtained by choosing the prior $g(\boldsymbol{\theta})=p(\boldsymbol{\theta}\,|\,M)$ as the importance distribution in Eq. 8:

$$p(\mathbf{D}\,|\,M)=\frac{1}{n}\sum_{i=1}^{n}p_{\beta_0}(\mathbf{D}\,|\,\boldsymbol{\theta}_i,M) \tag{9}$$

AM is subject to likelihood rounding error and not a acceptance ratio rounding error. In other words, we can accurately sample the target parameter values using MC simulation, but we are unable to correctly evaluate their likelihoods values due to rounding error. To reduce likelihood rounding error we use the log-likelihood and factor out largest likelihood $L_{\max} = \max_{1<i<n} p_{\beta_0}(\mathbf{D}|\boldsymbol{\theta}_i, M)$, yielding the AM estimation of BME as

$$\hat{r}_{AM} = \frac{1}{n}\exp(LogL_{\max})\sum_{i=1}^{n}\exp\left[Logp_{\beta_0}(\mathbf{D}\,|\,\boldsymbol{\theta}_i, M) - LogL_{\max}\right] \tag{10a}$$

and in the log form as

$$\log \hat{r}_{AM} = \log L_{\max} + \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(\log p_{\beta_0}(\mathbf{D}\,|\,\boldsymbol{\theta}_i, M) - \log L_{\max}\right)\right) \tag{10b}$$

given a sample size, where $p_0(\mathbf{D}\,|\,\boldsymbol{\theta}_i, M)$ is likelihood of sample $\boldsymbol{\theta}_i$ drawn from the prior parameter distribution $p(\boldsymbol{\theta}\,|\,M)$. Eq. 10 clearly shows that likelihood rounding error cannot be completely removed due to underflow. Any sample with $\exp\left[Logp_{\beta_0}(\mathbf{D}\,|\,\boldsymbol{\theta}_i, M) - LogL_{\max}\right]$ smaller than the smallest nonzero denormalized number ε_{dn} will be assigned a value of zero. Being subject to underflow, AM is a biased estimator in practice that will underestimate the BME regardless of the sample size n.

Alternatively, estimating the BME through evaluating the posterior using MCMC can be much more computationally efficient than through evaluating the prior using Monte Carol simulation. Using the posterior $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta}\,|\,\mathbf{D}, M)$ as the importance distribution in Eq.8 yields the harmonic mean estimator(Newton and Raftery 1994):

$$\hat{r}_{HM}(\mathbf{D}|M) \approx n \Big/ \sum_{i=1}^{n} \frac{1}{p_1(\mathbf{D}\,|\,\boldsymbol{\theta}_i, M)} \tag{11}$$

where $p_1(\mathbf{D}\,|\,\boldsymbol{\theta}_i, M)$ is likelihood of sample $\boldsymbol{\theta}_i$ drawn from the posterior parameter distribution $p_1(\boldsymbol{\theta}\,|\,\mathbf{D}, M)$. As there is no closed form expression for the posterior for most practical applications, we generally use MCMC to approximate and draw sample from the posterior. Although HM is susceptible to acceptance ratio rounding error, it is very unlikely to happen as posterior samples will generally have high likelihood. Accordingly, the posterior probability ratio $\alpha_{posterior}$ (Eq. 6c and Eq. 7) is not expected to suffer from subnormal underflow or overflow, when using likelihood function or the log-likelihood function, respectively.

The thermodynamic integration is based on technique of path sampling that is explained in Section 2.1. We are interested in computing the ratio in Eq. 5, which can give the log-BME according to TI as (Lartillot and Philippe 2006):

$$\log p(\mathbf{D}\,|\,M) = \log\left(\frac{Z_1}{Z_0}\right) = \int_0^1 \frac{\partial \log Z_\beta}{\partial \beta}\partial\beta = \int_0^1 E_\beta[\log p(\mathbf{D}\,|\,\boldsymbol{\theta}, M)]\partial\beta = \int_0^1 y_\beta \partial\beta \qquad (12)$$

such that the potential expectation y_β can be discretized by discrete power coefficients β. Using any quadrature rules, the one-dimensional integral in Eq. 12 can be solved by incrementally sampling from $\beta_0 = 0$ to $\beta_K = 1$, and taking the expectation of the log-likelihood $\log p_{\beta_k}(\mathbf{D}\,|\,\boldsymbol{\theta}, M)$ at each $\beta_k \in [0,1]$. Using the trapezoidal rule and taking the exponent, Eq. 12 can be solved as:

$$\hat{r}_{TI} \equiv p(\mathbf{D}\,|\,M) = \exp\left(\int_0^1 y_{\beta_k} d\beta\right) = \exp\left(\sum_{k=0}^{K} r_{TI,k}\right) \qquad (13a)$$

such that

$$r_{TI,k} = (\beta_k - \beta_{k-1})\left[\frac{y_k - y_{k-1}}{2}\right] \qquad (13b)$$

and

$$y_k = E_\beta[\log p(\mathbf{D} \mid \boldsymbol{\theta}_k, M)] = \frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{D}\left|\boldsymbol{\theta}_{k,i}, M\right.) \tag{13c}$$

TI converts the multidimensional integral (Eq. 4) to one dimensional integral (Eq. 13) with known β_k values in the *x*-axis and unknown y_k values in the y-axis. Each panel in this one-dimensional integral is given by $r_{TI,k}$ (Eq. 13b) for the case of trapezoidal rule. The summation of these panels gives the log-BME of TI (Eq.13a). A one practical note is that the reader might wonder that in the presence of subnormal underflow, working with likelihood function will create a problem in the post processing step of estimating the BME using Eq. 13. In other words, if the likelihood $p(\mathbf{D}\left|\boldsymbol{\theta}_{k,i}, M\right.)$ underflow to zero and was stored in the computer memory as zero, then $\log(p(\mathbf{D}\left|\boldsymbol{\theta}_{k,i}, M\right.)) = \log(0) = -\inf$ and a direct substitution of this to Eq. 11c gives $y_k = -\inf$. This should not be a problem because although we run MCMC with likelihood function, we can retain the residuals of the MCMC samples and recalculate their log-likelihoods $\log p(\mathbf{D}\left|\boldsymbol{\theta}_{k,i}, M\right.)$ for Eq.13c in the post processing step. This is similar to using log-likelihoods with AM (Eq. 10).

TI is not susceptible to likelihood rounding error but to acceptance ratio rounding error. Unlike importance sampling estimators, TI has a different approach that directly uses log-likelihood and thus is immune against likelihood rounding error. By comparing AM (Eq. 10) and TI (Eq. 13c), we can clearly see that TI is not subject to likelihood rounding error since we take the average of the log-likelihoods $\log p(\mathbf{D}\left|\boldsymbol{\theta}_{k,i}, M\right.)$. Unfortunately, TI is subject to acceptance ratio rounding error when using MCMC sampling. To estimate y_k (Eq. 13c) at each power posterior interval β_k, we sample $\boldsymbol{\theta}_{k,i} \sim p_{\beta_k}(\boldsymbol{\theta} \mid \mathbf{D}, M)$ directly using MC simulation if an analytical expression is available. However, for most practical applications the stationary distributions $p_{\beta_k}(\boldsymbol{\theta} \mid \mathbf{D}, M)$ are approximated and sampled using MCMC. The only exception is the first power

posterior density $p_{\beta_0}(\boldsymbol{\theta}\,|\,\mathbf{D},M)$ that is the prior for which drawing samples using MC simulation or MCMC gives similar results. Corresponding log-likelihoods are calculated and substituted in Eq.13c to obtain y_0 at the prior. The y_0 will not suffer from likelihood rounding error as we take the mean of the log-likelihoods $\log p(\mathbf{D}\left|\boldsymbol{\theta}_{k,i},M\right.)$ (Eq. 11c) and will not suffer from any acceptance ratio rounding error since all samples are accepted (Eq. 6 and Eq. 7). However, any other power posterior density $p_{\beta_k}(\boldsymbol{\theta}\,|\,\mathbf{D},M)$ that needs to be approximated and sampled using MCMC will be subject to subnormal underfollow as discussed by Eq. 6 when working with likelihood function, and underflow and overflow as discussed by Eq. 7 when working with log-likelihood function. Accordingly, sampling power posterior densities $p_{\beta_k}(\boldsymbol{\theta}\,|\,\mathbf{D},M)$ close to the prior will be hampered by acceptance ratio rounding error, resulting in overestimating y_k. Eventually, this overestimation will reduce as we move closer to the posterior. However, this overestimation of y_k is cumulative as indicated by Eq. 13a resulting in BME overestimation.

The importance sampling SS estimator borrows ideas from path sampling technique. Again, we are interested in computation the ratio in Eq. 5, which can be expressed as a product of K ratios:

$$p(\mathbf{D}\,|\,M)=\frac{Z_1}{Z_0}=\left(\frac{Z_1}{Z_{\beta_{K-1}}}\right)\left(\frac{Z_{\beta_{K-1}}}{Z_{\beta_{K-2}}}\right)\ldots\left(\frac{Z_{\beta_k}}{Z_{\beta_{k-1}}}\right)\left(\frac{Z_{\beta_{k-1}}}{Z_{\beta_{k-2}}}\right)\ldots\left(\frac{Z_{\beta_2}}{Z_{\beta_1}}\right)\left(\frac{Z_{\beta_1}}{Z_0}\right)=\prod_{k=1}^{K}\frac{Z_{\beta_k}}{Z_{\beta_{k-1}}} \tag{14}$$

As noted by Xie et al. (2011), $g(\boldsymbol{\theta})=p_{\beta_{k-1}}(\boldsymbol{\theta}\,|\,\mathbf{D},M)$ is an excellent importance distribution as it is only slightly dispersed than $p_{\beta_k}(\boldsymbol{\theta}\,|\,\mathbf{D},M)$. Thus, using the path defined in Eq. 1, each ratio $r_{SS,k}=Z_{\beta_k}/Z_{\beta_{k-1}}$ can be expressed as:

$$r_{SS,k}=\frac{Z_{\beta_k}}{Z_{\beta_{k-1}}}=\frac{\int q_{\beta_k}(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int q_{\beta_{k-1}}(\boldsymbol{\theta})d\boldsymbol{\theta}}=\frac{1}{n}\sum_{i=1}^{n}p(\mathbf{D}\left|\boldsymbol{\theta}_{k-1,i},M\right.)^{\beta_k-\beta_{k-1}} \tag{15}$$

in which $\boldsymbol{\theta}_{k-1,i} \sim p_{\beta_{k-1}}(\boldsymbol{\theta} \mid \mathbf{D}, M)$ is sampled from the power posterior density $p_{\beta_{k-1}}(\boldsymbol{\theta} \mid \mathbf{D}, M)$. By comparing Eq. 13 and Eq. 15 we note that for the last ratio $r_{TI,K}$ of TI we sample $p_{\beta_{k-1}}(\boldsymbol{\theta} \mid \mathbf{D}, M)$, while for the last ratio $r_{SS,K}$ of SS we sample $p_{\beta_{k-1}}(\boldsymbol{\theta} \mid \mathbf{D}, M)$. Thus, interestingly, SS does not need to sample the posterior $p_{\beta_{k-1}}(\boldsymbol{\theta} \mid \mathbf{D}, M)$. Finally, SS estimate the BME by combining all K ratios:

$$\hat{r}_{SS} \equiv p(\mathbf{D} \mid M) = \prod_{k=1}^{K} \frac{Z_{\beta_k}}{Z_{\beta_{k-1}}} = \prod_{k=1}^{K} r_{SS,k} \tag{16a}$$

and rewriting Eq. 15 to improve the numerical stability:

$$r_{SS,k} = \frac{1}{n} (L_{\max,k-1})^{\beta_k - \beta_{k-1}} \sum_{i=1}^{n} \exp\left\{ (\beta_k - \beta_{k-1}) \left[\log p(\mathbf{D} \mid \boldsymbol{\theta}_{k-1,i}, M) - \log L_{\max,k-1} \right] \right\} \tag{16b}$$

Factoring out the largest likelihood $L_{\max,k-1} = \max_{1<i<n} p(\mathbf{D} \mid \boldsymbol{\theta}_{k-1,i}, M)$ and substituting the log-likelihood values similar to AM (Eq.10) will reduce the numerical error. Note that for the special case of $K = 1$, the prior p_{β_0} becomes the only importance distribution, and thus SS (Eq.16) reduces to AM (Eq.10).

SS is less susceptible to these both rounding error artifacts. SS does not generally suffer from likelihood rounding error. Comparing AM (Eq. 10) and SS (Eq. 16b) implies that likelihood rounding error could occur when $\exp\left\{ (\beta_k - \beta_{k-1}) \left[\log p(\mathbf{D} \mid \boldsymbol{\theta}_{k-1,i}, M) - \log L_{\max,k} \right] \right\} < \varepsilon_{dn}$ resulting in an underflow. However, likelihood rounding error will not occur with SS due to the multiplication term $(\beta_k - \beta_{k-1})$. On the other hand, the formulation of SS (Eq. 18a) can potentially reduce the impact of acceptance ratio rounding error on BME estimation. The acceptance ratio rounding error discussed for TI are equally applicable for SS as we use the same MCMC samples for both estimators. However, by multiplying SS ratios $r_{SS,k} = Z_{\beta_k} / Z_{\beta_{k-1}}$ for successive distributions $p_{\beta_k}(\boldsymbol{\theta} \mid \mathbf{D}, M)$, the acceptance ratio rounding error is more or less a systematic error

that can be factored out. This can potentially but not necessarily reduce the impact of acceptance ratio rounding error as shown in the results section. On the other hand, by summing the $r_{TI,k}$ values for TI (Eq.13a), acceptance ratio rounding error will sum up.

2.4 Discretizing the path from the prior to the posterior

The point that was left undiscussed is optimum discretization of the path through β values selection to reduce the discretization error. Unlike likelihood rounding error and acceptance ratio rounding error, discretization error can be reduced (and almost eliminated) by investing in the computational effort. TI and SS introduce an approximation by discretization of the path joining the prior and posterior. The remaining question is how to choose the schedule of the power coefficients β to minimize this discretization error.

The discretization error is a function of the number and values of β_k. Obviously increasing the number K of the power posterior distributions $p_{\beta_{k-1}}(\boldsymbol{\theta}\,|\,\mathbf{D},M)$ will reduce the discretization error. As the discretization error depends on the Kullback-Liebler distance between successive power posterior distributions $p_{\beta_{k-1}}(\boldsymbol{\theta}\,|\,\mathbf{D},M)$, the computational efficiency can be increased by refining the discretization not along the whole path but along the segments where the Kullback-Liebler distance between successive distributions is large (Calderhead and Girolami 2009; Friel et al. 2014; Lefebvre et al. 2010). In practice it is often desirable to place more β values where the power posterior is changing rapidly (Calderhead and Girolami 2009; Friel and Pettitt 2008; Xie et al. 2011). This can be done by adaptive choice of β values (Friel et al. 2014), manual iterative construction of β schedule (Liu et al. 2016), or using an algebraic function (Friel and Pettitt 2008; Lartillot and Philippe 2006; Xie et al. 2011). Following Xie et al. (2011) we use the positively skewed Beta(α,1) distribution to assign the β values such that

$$\beta_k = (k/K)^{1/\alpha} \tag{17}$$

for $k = 1,2,...,K$, and $\beta_0 = 0$. For $\alpha = 1$ the β values are equally spaced similar to Lartillot and Philippe [2006]. As power posterior is changing rapidly near the prior, setting $\alpha = 0.25$ places most the β values near the prior $\beta_0 = 0$ (Friel and Pettitt 2008). Xie et al. (2011) suggest that $\alpha = 0.3$ is a reasonable choice such that half of β values is less than 0.1. In this study we use Eq. 17 with $\alpha = 0.3$ to generate the power posterior coefficients β schedule, unless otherwise stated.

3. Likelihood rounding error

3.1 A high dimensional Gaussian model

We show the degree to which AM, HM, TI and SS are sensitive to likelihood rounding error using a Gaussian model, because the true BME is available analytically and we can make direct draws from the power posteriors using MC simulation. Thus, by avoiding errors from MCMC approximation, we can separately look at the likelihood rounding error. The main purposes of this exercise are (i) to show the impact of likelihood rounding error on the four estimators, (ii) to show how each estimator evades this problem differently, and (iii) to discuss the implications of our findings of to BME estimation. We use a D-dimensions Gaussian model parameterized by $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_D)$ such that the prior is the product of independent standard normal distribution of each parameter θ_d; the likelihood is the product of:

$$p(M \mid \boldsymbol{\theta}) = \prod_{d=1}^{D} e^{-\frac{\theta_d^2}{2v}} \tag{18}$$

where v represents the variance; and samples for all power posterior distributions p_{β_K} can be directly drawn from a normal distribution (Lartillot and Philippe 2006):

$$\theta_{d,\beta_k} \sim N\left(0, \sqrt{\frac{v}{\beta_k + v}}\right) \tag{19}$$

with mean 0 and variance $v/(\beta_k + v)$. The true analytical solution (AS) of the BME is

$$\log p(M) = \frac{D}{2}\log\left(\frac{v}{1+v}\right) \tag{20}$$

for a model with D dimensions.

For a 100-dimensional problem, we parameterize the path $\beta_0 = 0$ to $\beta_K = 1$ using Eq. 14 with $K = 100$ and directly draw $n = 20000$ samples for each β_k step. With a total number of $N = n(K+1)$ samples, the AM, HM and TI estimator directly draw samples from prior, posterior and power posterior paths, respectively. SS needs only $N = n(K)$ samples since SS does not require sampling the posterior. We test three cases with different variances $v = 1$, 0.1, and 0.01. For variance $v = 1$ the likelihood function is relatively flat. As the variance decreases, the likelihood function becomes more peaked and concentrated such that for the case of $v = 0.01$ the BME becomes very small due to the presence of large parameter space with small likelihoods. Table 2 reports the mean log-BME based on 10 independent estimations and the true analytical solution. For the case of $v = 1$, the performance of the four methods is relatively good, with AM and HM showing a slight underestimation and overestimation, respectively. Yet by decreasing the variance v such that the likelihood function becomes more peaked and concentrated, the AM and HM estimations substantially deteriorated while TI and SS maintained their accurate estimations.

3.2 Impact of likelihood rounding error

If these four estimators use samples from the same distributions, then why is AM prone to suffer most from likelihood rounding error, while HM, TI and SS are not? Prior AM sampling involves drawing samples with extremely small likelihood. To address this issue, the AM strategy

of using log-likelihoods, and factoring out the largest log-likelihood (Eq.10) is not very effective. Likelihood rounding error caused AM to underestimate the BME by orders of magnitude for variance $v = 0.01$ as shown in Table 2. By looking at AM estimator (Eq.10), the likelihood term $\exp\left[Logp_{\beta_0}(\mathbf{D} \mid \boldsymbol{\theta}_i, M) - LogL_{\max} \right]$ underflows to zero for about 99.84% of the samples for the case of $v = 0.01$. While the prior sampling of AM estimator makes it most vulnerable to likelihood rounding error, the posterior HM estimator (Eq. 11) did not experience any likelihood rounding error. None of the likelihoods $Logp_{\beta_0}(\mathbf{D} \mid \boldsymbol{\theta}_i, M)$ of the posterior samples underflow. Yet Table 2 shows that the HM estimations can be very biased since the posterior samples have high likelihoods, which is discussed letter. Interestingly, TI and SS experience the two limiting cases of AM and HM as a path is sampled from the prior to the posterior, yet the estimations of TI and SS for the three variance cases are not biased as shown in Table 2. None of the TI samples underflow to zero. This is because TI avoids taking the exponent of the log-likelihood like AM (Eq.10), but directly take the mean of the log-likelihood (Eq.13c). For SS, although the likelihood terms $\left[\log p(\mathbf{D} \mid \boldsymbol{\theta}_{k-1,i}, M) - \log L_{\max,k} \right]$ of AM and SS are the same, none of the SS samples underflow to zero due to the multiplication term $\left(\beta_k - \beta_{k-1} \right)$ as previously explained (Eq.16).

Table 2. For the 100-dimensional Gaussian model, true analytical solution (AS) of the log-BME with different variance (v), and the corresponding numerical estimates of arithmetic mean (AM), harmonic mean (HM), thermodynamic integration (TI), steppingstone sampling (SS), using $N = n(K+1)$ samples such that $K = 100$ and $n = 2000$. The numerical estimates are the mean log-BME (±standard deviation) of 10 independent estimations.

v	AS	AM	HM	TI	SS
1	-34.66	-34.89 (±0.53)	-33.49 (±0.55)	-34.66 (±0.01)	-34.66 (±0.01)
0.1	-119.89	-254.78 (±9.68)	-69.26 (±1.69)	-119.91 (±0.04)	-119.89 (±0.04)
0.01	-230.76	-2437.24 (±108.2)	-75.99 (±1.86)	-230.87 (±0.07)	-230.76 (±0.07)

There are two implications for these findings. First, while it is generally assumed that AM estimator slowly yet eventually approximates the true solution as the sample size increases (Liu et

al. 2016; Schöniger et al. 2014), we show that, in the presence of likelihood rounding error, the AM estimator never converges to the true solution. The reasons for the AM underestimation are different for both cases of variance $v = 0.1$ and variance $v = 0.01$ shown in Table2. For the case of $v = 0.01$, the likelihood of more than 99% of the samples underflow to zero, while for variance $v = 0.1$ none of the samples underflows to zero. Obviously, AM underestimated the BME for the case of $v = 0.01$ due to numerical underflow. For the case of $v = 0.1$, AM underestimated the BME, because the probability of drawing a sample with high likelihood is extremely small and most samples generated from the prior space have small likelihood (Liu et al. 2016; Schöniger et al. 2014). This is not the case for $v = 1$ because its likelihood function is not narrow. Generally, the narrower the likelihood function, the more sampling effort is needed for the AM to converge to the true solution. In Table 3 we additional show that convergence of AM can be extremely slow. For example, for the case of $v = 0.1$ with one billion samples the AM estimation of log-BME is -198.44 (±5. 81), while the true analytical solution of the log-BME is -119.89. More critically, with a much peaked and contracted likelihood function as the case of $v = 0.01$ in which likelihood rounding error will start to step-in, AM will never converge to the true solution regardless of the computational effort. As shown in Table 3 for the case of $v = 0.01$, with more sampling effort the AM bias is very slowly decreasing. Using one billion samples, the AM log-BME estimation of -1890.12 (±70.54) is biased by orders of magnitude from the true analytical solution of -230.76. As the sample size goes to infinity, this huge bias due likelihood rounding error will not converge to the true solution, but to the smallest nonzero denormalized floating-point number $\log(\varepsilon_{dn}) = -745.13$.

Table 3. For the 100-dimensional Gaussian model, log-BME bias of arithmetic mean (AM) and harmonic mean (HM) estimations with different sample sizes. The bias is the difference between estimated solution and the true solution. For the two cases of variance v=0.1 and variance v=0.01, the true analytical solutions are -119.89 and -230.76, respectively. We report the mean (±standard deviation) of log-BME bias given 10 independent estimations.

Sample size	v=0.1		v=0.01	
	AM	HM	AM	HM
1×10^3	-185.87 (±20.14)	57.15 (±2.47)	-2799.67 (±147.64)	163.84 (±1.42)
1×10^4	-161.65 (±12.25)	55.97 (±2.38)	-2521.14 (±130.22)	160.41 (±1.76)
1×10^5	-142.88 (±7.65)	51.42 (±2.68)	-2282.33 (±109.11)	155.72 (±1.27)
1×10^6	-119.55 (±7.21)	49.81 (±1.19)	-2057.36 (±72.69)	152.43 (±2.94)
1×10^7	-99.57 (±9.25)	45.36 (±3.22)	-1872.51 (±73.01)	149.14 (±3.15)
1×10^8	-89.82 (±7.56)	44.56 (±2.01)	-1740.55 (±71.24)	147.72 (±2.19)
1×10^9	-78.55 (±5. 81)	42.51 (±1.14)	-1659.36 (±70.54)	145.45 (±2.12)

The second implication is for HM. One might argue that, since HM is free from likelihood rounding error, HM can eventually converges to the true solution. This is true in theory but not in practice. The posterior samples have high likelihoods, which results in larger overestimation as the likelihood function keeps narrowing. For example, Table 2 shows that while the HM estimation for variance $v = 1$ is relatively accurate, HM overestimated the BME by orders of magnitude for variance $v = 0.01$. This is because when we tighten the variance $v = 0.01$ we are stating that a very narrow region of the parameter space is most likely, and the probability of drawing a sample with a small likelihood becomes very low. As we narrow the likelihood function by reducing the variance, it become more difficult for the HM to converge. In practice, increasing the sample size will be of much help to HM as shown by Table 3. For example, using one billion samples the bias of the HM estimations for variance $v = 0.1$ and variance $v = 0.01$, is 42.51 (±1.14) log units and 145.45 (±2.12) log units, respectively. Thus, although HM does not suffer from likelihood rounding error, it suffers from slow convergence as the likelihood function becomes narrower.

We are now left with the TI and SS estimators, which passed both tests of likelihood rounding error and slow convergence. While BME is by definition the weight average of the likelihood functions where weights from the prior, the AM estimator is indeed the simplest

mathematical expression for BME. However, the AM estimator is most susceptible to underflow of the likelihoods due to rounding error (i.e., likelihood rounding error). Although they are merely elaborate mathematical expressions for solving the same problem, TI and SS can indeed avoid rounding error. It is interesting to see that we can avoid rounding error through methodological development. However, for most practical applications, TI and SS loss this relative advantage due to the use of MCMC since we cannot avoid rounding error with the Metropolis acceptance ratio (i.e., acceptance ratio rounding error) as shown in the following section.

4. Sampling rounding error

4.1 A linear two-dimensional model

A simple two dimensional linear test function suffices to illustrate the acceptance ratio rounding error. We use the liner model M described in the study of Lu et al. (2012):

$$\mathbf{y} = a\mathbf{x} + m + \boldsymbol{\varepsilon} \tag{21}$$

with true parameter $\boldsymbol{\theta} = \{a, m\}$ of $a = 2$ and $m = 3$. A single data set $\mathbf{D}$ of size 20 was generated by corrupting the y values corresponding to $\mathbf{x} = \{1,...,20\}$ with a realization of random error $\boldsymbol{\varepsilon} = \{\varepsilon_1,...,\varepsilon_{20}\}$ drawn from a standard normal distribution $\varepsilon \sim N(0,1)$. The simulated data set is available in the Supplementary Material. Given a Gaussian likelihood function and the conjugate priors of $a \sim N(2,1)$ and $m \sim N(3,1)$, an exact analytical solution of the BME $p(\mathbf{D} \mid M) = 5.1194 \times 10^{-14}$ is obtained using Mathematica, and is used as the reference solution.

Estimating BME using TI and SS is subject to the error sources listed in Table 1. From the previous analysis we know that TI and SS relatively immune to likelihood rounding error. We try to remove MCMC sampling error by both using the state of the art MCMC code of MT-DREAM$_{(ZS)}$ (Laloy and Vrugt 2012), and choosing a simple two dimensional linear test function. The main purposes of this example are (i) to show how acceptance ratio rounding error can

dominate stochastic sampling error and discretization error when using TI, (ii) to show the impact of acceptance ratio rounding error on MCMC sampling with both likelihood function and log-likelihood function, (iii) to explain how TI does not pass the test of acceptance ratio rounding error, and (iv) to explain why SS can potentially better handle acceptance ratio rounding error than TI.

4.2 Dominance of acceptance ratio rounding error

This section shows that for TI the likelihood rounding error dominates over other error sources. As we except likelihood rounding error and MCMC sampling error to be minimal, we focus on the domination of acceptance ratio rounding error over MC stochastic error and discretization error. Except at β_0 , we use MCMC to draw $n = 100000$ samples at each β_k and burn-in the first 1000 samples. Note that for MCMC sampling, although we work with standard least square likelihood function, we retain the residuals of accepted samples to calculate their log-likelihoods. In a post-analysis to sample drawing the log-likelihoods are substituted in TI estimator (Eq. 13) and SS estimator (Eq. 16) to compute the BME. For the two cases of coarse discretization $K = 10$ and fine discretization $K = 200$ of the sampling path, Figure 1 shows the BME estimations of both TI and SS at sequence of n intervals. We report the mean and standard deviation of the BME of 10 independent MCMC runs. For the case of $K = 10$, Figure 1a shows that increasing the sample size n eventually decreases the solution dispersion of repeated runs, and that the mean estimation bias eventually converges such that MC stochastic error becomes very small for both TI and SS. While finer discretization should reduce the estimation bias, Figure 1b shows that TI estimation has a large systematic bias for the case of $K = 200$. Given that finer discretization by using more K should reduce the estimation bias, such counter intuitive results are due to acceptance ratio rounding error. We explain this in details in the next section. Moreover, Figure 1b shows that this large systematic bias of TI is not reduced by increasing sample size n. In summary,

in the presence of acceptance ratio rounding error the estimation bias of TI does not improve by increasing the sample size n (i.e., reducing the MC stochastic error) or increasing the number of steps K (i.e., reducing the discretization error).

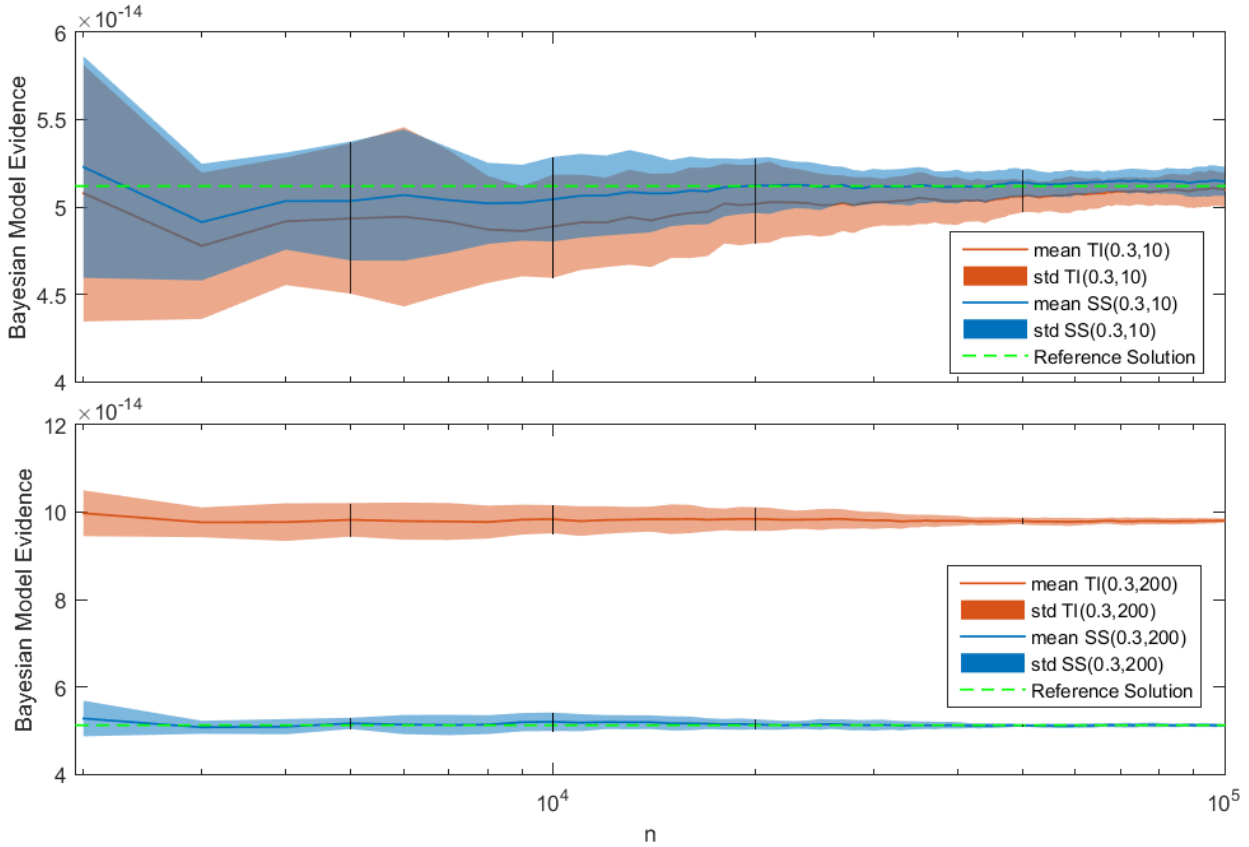

Figure 1. For the two-dimensional linear model, by increasing the sample size n at each power posterior coefficient β_k, we report the BME (mean and standard deviation) of 10 independent estimations using thermodynamic integration $\mathrm{TI}(\alpha, K)$ and steppingstone sampling $\mathrm{SS}(\alpha, K)$ for $K = 10$ and 200 in the top and bottom plots, respectively. The number of power posterior coefficients β are $K+1$ and their values are determined by Eq. 17 according to the shape parameter α. For each independent estimation, the total number of samples are $N = n(K+1)$.

In the presence of rounding error, the estimation bias of TI does not depend on the number of power poster coefficients K but on their values β_k. To investigate how SS and TI are sensitive to β_k values, we use Eq. 17 to generate three different paths using $\alpha = \{0.25,\ 0.3,\ 1.0\}$ with both coarse discretization $K = 10$ and fine discretization $K = 200$. Note that, for all the following

results, we use a simple size $n = 5000$ at each β_k and burn-in the first 1000 samples. Based on 10 independent MCMC runs, the bias and dispersion of BME estimations of both TI and SS are shown in Figure 2. The results show that the solution dispersion for repeated runs appears to improve for both TI and SS with increasing K. More importantly Figure 2 clearly shows that the bias of SS is relatively invariant to β_k values, while the bias of TI considerably depends on β_k values. The TI estimations seems erratic. For example, given the coarse discretization $K = 10$, the TI mean estimation for the three cases of $\mathrm{TI}_{(0.25,10)}$, $\mathrm{TI}_{(0.3,10)}$ and $\mathrm{TI}_{(1,10)}$, is overestimated, accurately estimated and underestimated, respectively. Note we use the notation of $\mathrm{Estimator}_{(\alpha,K)}$ for the remaining discussion. When using a fine discretization $K = 200$, TI bias even increased. It is unclear why the case of $\mathrm{TI}_{(0.3,10)}$ is unbiased while the case of $\mathrm{TI}_{(0.3,200)}$ is very biased, showing a relative error of -0.55% and 91.58%, respectively. Relative error is $\delta = [p_{est}(D \mid M) - p_{ref}(D \mid M)] / p_{ref}(D \mid M)$ in which $p_{ref}(D \mid M)$ and $p_{est}(D \mid M)$ are the reference solution and estimated solution of the BME, respectively. In practice, in the absence of a true reference solution, such apparently erratic results can be very frustrating. In this section we explain why this happened. To understand why TI gives such biased estimates, we show step-by-step how likelihood rounding error impacts the BME estimation for both cases of working with likelihood function and log-likelihood function during the MCMC sampling.

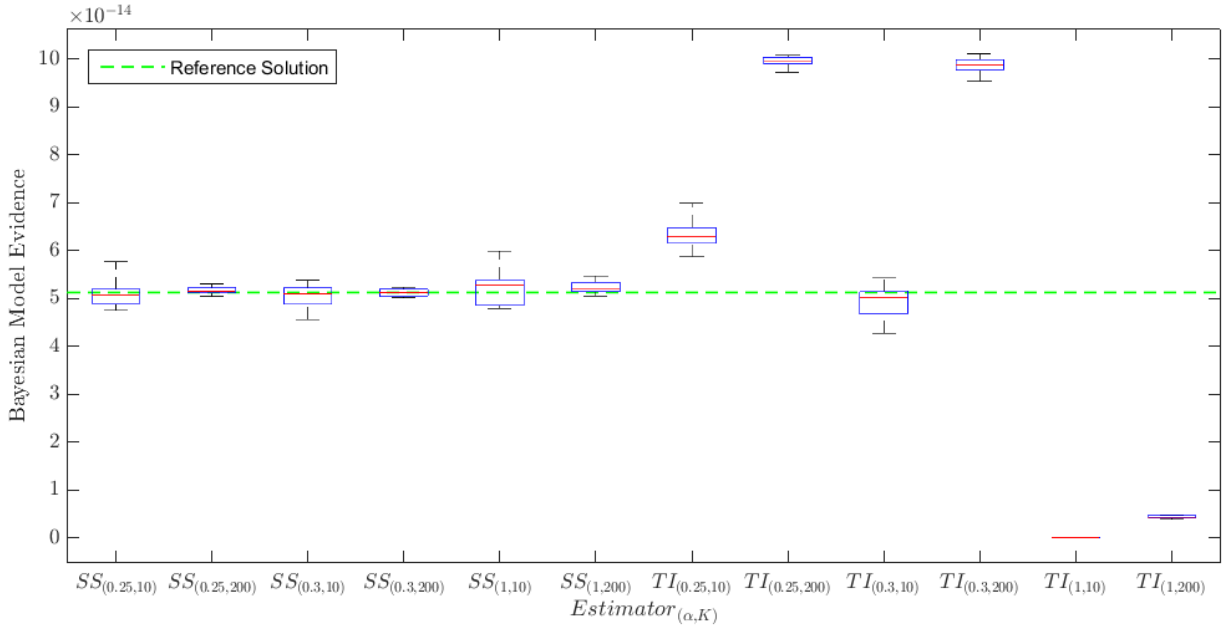

Figure 2. For the two-dimensional linear model, BME estimations of steppingstone sampling (SS) and thermodynamic integration (TI) for three shape parameters $\alpha = \{0.25, 0.3, 1\}$ with small $K = 10$ and large $K = 200$. The number of power posterior coefficients β are $K+1$, and their values are determined by Eq.17 according to the shape parameter α. The results are based on 10 independent estimations. For each independent estimation, the total number of samples are $N = n(K+1)$ in which $n = 5000$.

4.3 MCMC sampling with likelihood function

We start by plotting the MCMC parameters for different β_k value for the three cases of $\mathrm{TI}_{(1,200)}$, $\mathrm{TI}_{(0.3,200)}$ and $\mathrm{TI}_{(0.3,10)}$ as shown in Figure 3. The figure shows that, for the three cases, the posterior space ($\beta_K = 1$) is significantly narrower than the prior space ($\beta_0 = 0$). The parameter samples of different β_k gradually spread near the posterior, yet then a huge jump occurs between the distributions of β_0 and β_1 for the three cases. In Figure 3b even the distributions of $\beta_1 = 2\times10^{-8}$ and $\beta_{20} = 0.0005$ look almost identical, although their β_k coefficients are four orders of magnitude apart. Is this true or $\beta_1 = 2\times10^{-8}$ distribution is inaccurately sampled? Given this wide prior, it is unlikely that the two distributions with four orders of magnitude apart be similar. On the contrary, the parameter samples of different β_k should gradually spread from the

posterior out to the prior space.

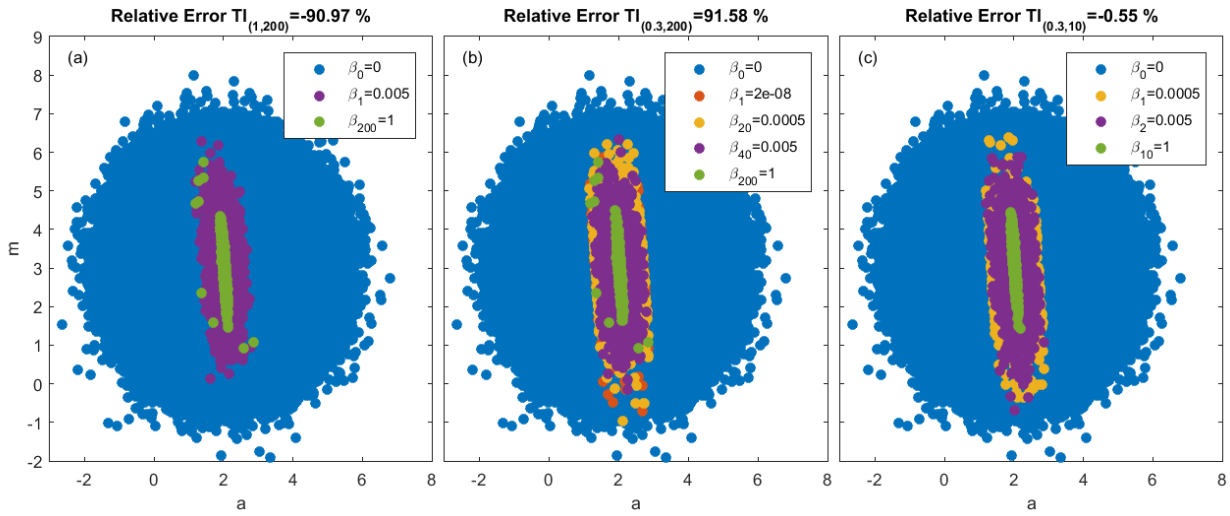

Figure 3. MCMC parameter $\boldsymbol{\theta} = \{a, m\}$ samples for different β_k values (Eq. 17) for the three cases of thermodynamic integration $\mathrm{TI}_{(\alpha,K)}$ for the two-dimensional linear model. The number of power posterior coefficients β_k are $K+1$, and their values are determined by Eq.17 according to the shape parameter α. We also report the relative error [%] of the BME estimation for each case.

For the case of $\mathrm{TI}_{(0.3,200)}$, the $\beta_1 = 2\times10^{-8}$ distribution is inaccurately sampled due to rounding error, resulting in BME overestimation. We are confident that the prior ($\beta_0 = 0$) is accurately sampled because it is directly sampled using MC, yet all other distributions are approximated and sampled using MCMC. For the $\beta_1 = 2\times10^{-8}$ distribution, the trace plot of the MCMC cycles (not shown here) indicates that large part of the parameter space is trimmed due to rounding error. The TI estimation is calculated by integrating potential expectations y_k from the prior ($\beta_0 = 0$) to the posterior ($\beta_{200} = 1$); the potential expectations $y_k = E[\log p_{\beta_k}(\mathbf{D}\,|\,\boldsymbol{\theta}_k, M)]$ of each β_k distribution is the expectation of the log-likelihoods of that distribution (Eq. 13c). All the parameter samples with likelihoods smaller than subnormal floating point number are trimmed (for reasons previously explained in the discussion of Eq. 6). Accordingly, the potential expectation

y_1 of $\beta_1 = 2 \times 10^{-8}$ is overestimated. This will add an artifact corner to the TI integral resulting in BME overestimation. Thus, the BME estimation of case $\mathrm{TI}_{(0.3,200)}$ has a relative error of 91.58%.

In the presence of acceptance ratio rounding error, assigning a value for β_1 seems to be the problem that caused TI to give bias estimations. In Figure 3a, assigning a large value for $\beta_1 = 0.005$ causes BME underestimation (as important corners in the integral is missing). In Figure 3b, assigning a very small value for $\beta_1 = 2 \times 10^{-8}$ caused BME overestimation (as additional corners are added to the integral due to rounding error). In Figure 3c, assigning an in between value for $\beta_1 = 0.0005$ gives unbiased estimation even with only $K = 10$ (as the integral is correct). To understand this reasoning, Figure 4 plots TI integral of y_k with β_k (Eq. 13a) for these three cases shown in Figure 3. As shown in Figure 4a, the full integrals from the prior ($\beta_0 = 0$) to the posterior ($\beta_K = 1$) of the three cases look the same since the potential expectations y_k changes dramatically near the prior. Thus, we need to zoom at this important integral interval near the prior $\beta \in [0, 0.01]$ where the power posterior distributions are changing rapidly (Figure 4b). Assuming the integral of $\mathrm{TI}_{(0.3,10)}$ to be more or less close to the unknown true integral as it has a small relative error of -0.55%, Figure 4b shows that $\mathrm{TI}_{(1,200)}$ missed the corner at $\beta = 4.6 \times 10^{-3}$ resulting in underestimating the BME by a relative error of -90.96%. Figure 4b also shows that power posterior is changing rapidly for $\beta < 4.6 \times 10^{-3}$ (i.e., y_k changing from -1461 to -246). The rule of thumb is to put more β_k values where the power posterior is changing rapidly, yet we cannot correctly estimate y_k for all these distributions with $\beta < 4.6 \times 10^{-3}$ (except at the prior $\beta_0 = 0$) due to acceptance ratio rounding error. We illustrate this by further zooming at the integral interval $\beta \in [0, 0.001]$ where we expect the acceptance ratio rounding error to take place as shown in Figure

4c. Again taking the integral of $\mathrm{TI}_{(0.3,10)}$ as our reference solution, for $\mathrm{TI}_{(0.3,200)}$ at $\beta_1 = 2\times10^{-8}$ we get $y_1 = -247$ while the reference solution is around $y = -1460$. Figure 4c clearly show that for $\beta < 4.7\times10^{-3}$ all the y_k values of $\mathrm{TI}_{(0.3,200)}$ are significantly overestimated, creating a nonexistent corner at $\beta_1 = 2\times10^{-8}$ as compared to the reference integral of $\mathrm{TI}_{(0.3,10)}$.

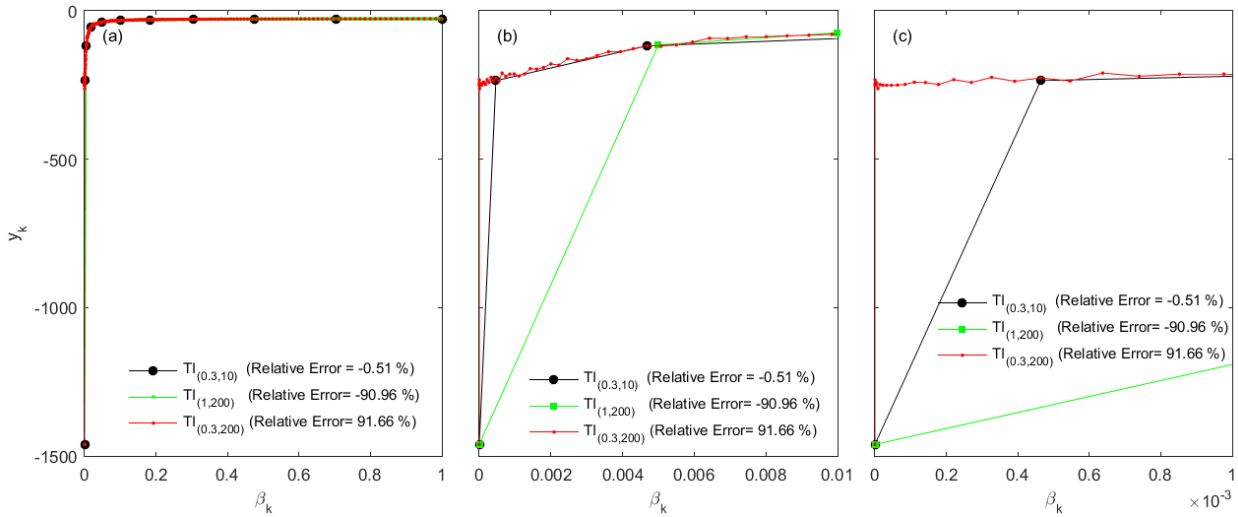

Figure 4. For the two-dimensional linear model, the thermodynamic integration (TI) integral (Eq. 13) shows the variation of the potential expectation y_k (Eq. 13c) with power posterior coefficient β_k for three cases $\mathrm{TI}_{(\alpha,K)}$. The number of power posterior coefficients β_k are $K+1$, and their values are determined by Eq.17 according to the shape parameter α. We also report the relative error [%] of the BME estimation for each case. Figure 4a shows the full integral from the prior ($\beta_0 = 0$) to the posterior ($\beta_K = 1$), while Figures 4b and 4c zoom at intervals close and closer to the prior ($\beta_0 = 0$), respectively.

Shifting the β_1 value away from the prior ($\beta_0 = 0$) can remove the systematic bias of TI. By considering the integral of $\mathrm{TI}_{(0.3,10)}$ to be our reference solution, the results show that for the case of $\mathrm{TI}_{(0.3,200)}$ the y_k values from $\beta_1 = 2.14\times10^{-8}$ to $\beta_{25} = 9.77\times10^{-4}$ are overestimated (Figure 4c). To confirm this observation, we remove these problematic values from the estimation such that $\beta_{26} = 1.11\times10^{-4}$ becomes the new β_1 instead of $\beta_1 = 2.14\times10^{-8}$ (and the corresponding BME estimation is based on $K = 175$ distributions instead of $K = 200$). This change drops the relative

error from $92.63 \pm 3.38\%$ to $-2.86 \pm 2.29\%$, based on 10 repeated runs. On the other hand, removing these problematic points $2.14 \times 10^{-8} < \beta < 9.77 \times 10^{-4}$ for $SS_{(0.3,200)}$ changed the relative error from $-0.03 \pm 1.49\%$ to $0.07 \pm 1.57\%$. It is clear that removing the inaccurately estimated y_k values will significantly improve TI, while SS remains relatively invariant as it performs well in the presence of acceptance ratio rounding error. It is worth noting that the results also suggest that SS is generally more accurate (in terms of bias and dispersion) than TI. This point is further supported by the relative errors of $\mathrm{TI}_{(0.3,10)}$ and $SS_{(0.3,10)}$ that are $-3.45 \pm 6.76\%$ and $-1.59 \pm 4.76\%$, respectively, based on 10 repeated runs.

In summary, we need to shift β_1 away from the prior to avoid acceptance ratio rounding error that causes BME overestimation. Yet as previously shown by the case of $\mathrm{TI}_{(1,200)}$ (Figure 3b and Figure 4b) shifting it "a little further" results in missing important corners in the integral, which causes BME underestimation. This is a deadlock, which we will inevitably encounter when working with likelihood function. However, one might argue that working with log-likelihood function is often preferred for its numerical stability. The theoretical analysis in Section 2.1 shows that this is only true when using MCMC to sample the posterior or power posterior distributions close to the posterior in the Kullback-Liebler sense. Rather, it is very problematic to work with log-likelihood to draw MCMC sample from posterior distributions near the prior distribution ($\beta_0 = 0$), which is demonstrated numerically in the next section.

4.4 MCMC sampling with log-likelihood function

One might argue that working with a log-likelihood function will avoid the subnormal underflow problem of the likelihood function. This is because the underflow limit of likelihood function (i.e., log of the smallest nonzero denormalized number $\log(\varepsilon_{dn}) = -744.44$) is extremely

large compared to that of log-likelihood function (i.e., negative real maximum $-R_{\max} \approx -1.7976 \times 10^{308}$). Thus, working with log-likelihood function permits sampling a wider prior space. For the two cases of β_1 and β_{25}, Figure 5 shows that indeed we can obtain wider parameter distributions when working with log-likelihood function as we can accept samples with likelihoods that are much smaller than $\log(\varepsilon_{dn}) = -744.44$. Accordingly, working with log-likelihood reduced the estimation bias. For example, for the previous case of $\mathrm{TI}_{(0.3,200)}$ the relative errors of the BME estimation for working with likelihood function and log-likelihood function are $92.63 \pm 3.38\%$ and $75.65 \pm 2.28\%$, respectively, based on 10 repeated MCMC runs. Yet the problem is not solved and the error is still as large as about 75%.

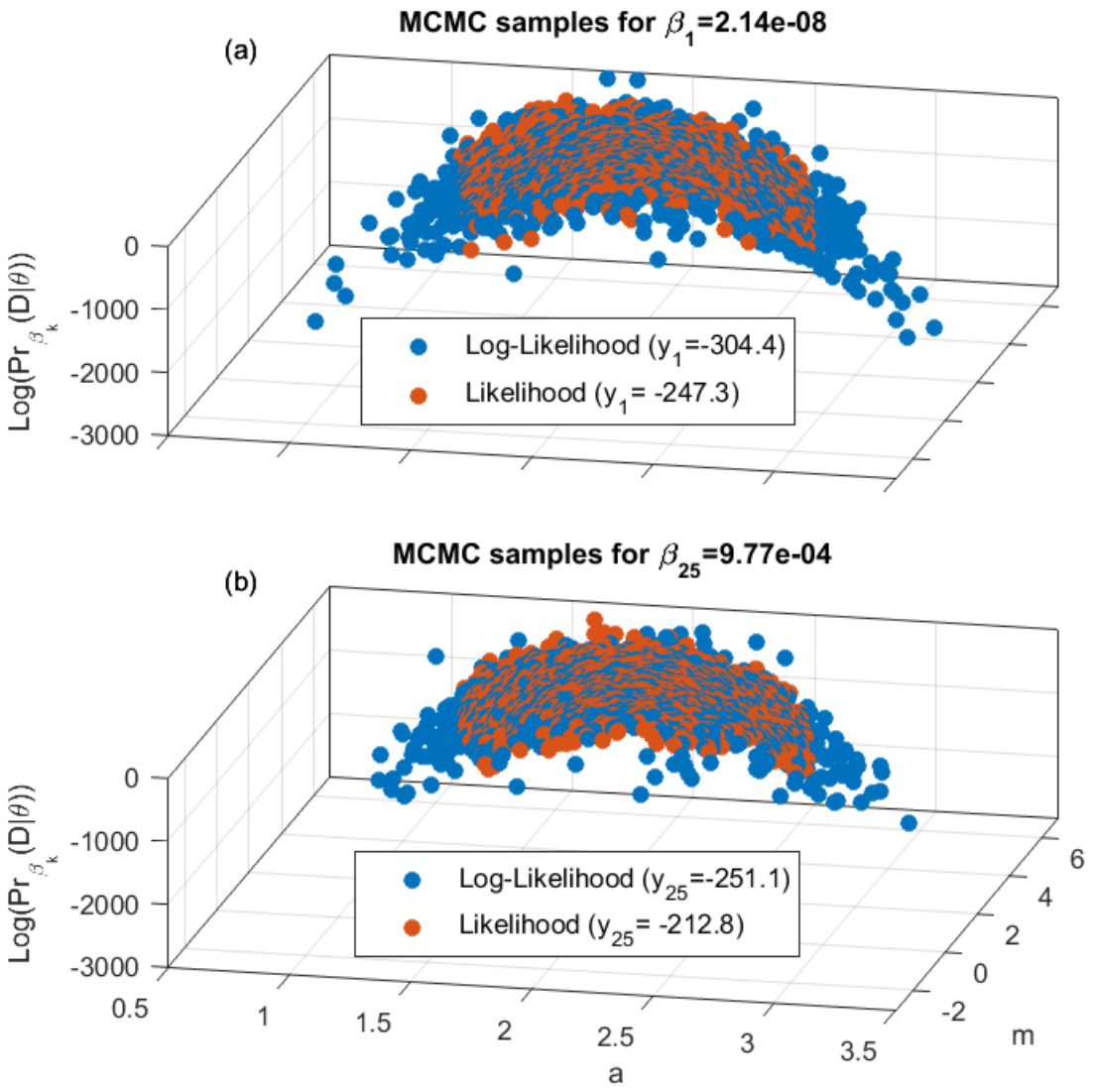

Figure 5. MCMC parameter samples $\boldsymbol{\theta} = \{a, m\}$ and their corresponding log-likelihoods are for the two-dimensional linear model. We show both cases of drawing MCMC samples using log-likelihood function and likelihood function for two power posterior distributions at (a) β_1 and (b) β_{25}. We also report the potential expectation y_k (Eq. 13c). Note the reference solutions for y_1 and y_{25} are -1461 and -218, respectively.

Actually, this slight reduction in the TI estimation bias when working with the log-likelihood function turned out to be a catch. From the previous discussion of Figure 4c, we obtained a reference solution of the potential expectation $y_1 = -1460$ at $\beta_1 = 2 \times 10^{-8}$. Working with log-likelihood function yields a potential expectation $y_1 = -304.3$ at $\beta_1 = 2 \times 10^{-8}$. Although

this is smaller than that of the likelihood function (i.e., $y_1 = -247.3$), it is still much larger than the reference solution (i.e., $y = -1460$). This indicates that we still cannot gradually move from β_0 to β_1 for very small β_k values even when working with log-likelihood function. Actually, this slight reduction in the potential expectation turned out to be a hidden sampling problem. At larger β_k this sampling problem yields y_k values that are even smaller than the reference solution. For example, at β_{25} the potential expectations for working with log-likelihood function and likelihood function are -251 and -212.8, respectively, while the reference solution is $y_{25} = -218.1$ (as computed from the results in Figure 4c). This sampling problem that appears when working with log-likelihood even affects the SS. Working with log-likelihood function increased the relative error of the SS estimation to $76.32 \pm 2.26\%$, versus $-0.03 \pm 1.49\%$ for working with likelihood function. This is mainly because log-likelihood has two sampling problems arising from acceptance ratio rounding error. Similar to working with likelihood function, the first problem is a numerical threshold that hinders sampling low likelihood regions in the parameter space (see the discussion of Eq. 6 for more details). A second and more serious problem is that working with log-likelihood function can corrupt the sampling ratios.

Acceptance ratio rounding error can corrupt the Metropolis acceptance ratio (Eq. 7) of MCMC when working with log-likelihood function. Table 4 shows the terms of the Metropolis acceptance ratio for four MCMC cycles from the stationary distribution of the power posterior distribution at $\beta_1 = 2.13 \times 10^{-8}$ for working with log-likelihood function. This distribution is wide enough since it is slightly less dispersed than the prior. Due to arithmetic overflow, the first cycle shows a case of accepting a sample that should have been rejected. The new sample should have been rejected due to its very small prior probability ratio $\alpha_{prior} = 3.97 \times 10^{-25}$. However, since

$\left[\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new},M)-\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new},M)\right]$ is larger than the log real maximum $\log(R_{\max})\approx 709.7827$, we obtained a posterior probability ratio $\alpha_{posterior}=Inf$, and the sample was unduly accepted. Thus, this numerical threshold invalidly widen the Metropolis acceptance ratio. This arithmetic overflow only appears when working with log-likelihood. The second cycle shows no bias as $\left[\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new},M)-\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new},M)\right]$ is less than the second threshold of smallest nonzero denormalized floating-point number $\log(\varepsilon_{dn})=-745.1332$. The third cycle shows a case of rejecting a sample that should have been accepted. This is due to the threshold imposed by smallest nonzero denormalized floating-point number, which results in underflowing to zero. This second threshold is also encountered when working with likelihood function. In the fourth cycle there is no bias as $\left[\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new},M)-\log(\mathbf{D}\,|\,\boldsymbol{\theta}_{new},M)\right]$ is smaller than the first threshed $\log(R_{\max})\approx 709.7827$ and the sample was correctly accepted. From the abovementioned discussion we conclude that when working with log-likelihood the MCMC cycles are confined by an upper threshold of $\log(R_{\max})\approx 709.7827$ resulting in overflowing to infinity and a lower threshold of $\log(\varepsilon_{dn})=-745.1332$ resulting in underflowing to zero. Thus, when working with log-likelihood function, these two limits act as systematic bias that shall control the MCMC sampling for any dispersed distribution.

Table 4. For the two-dimensional linear model, the old and new parameter samples $\boldsymbol{\theta} = \{a, m\}$ with their corresponding log-likelihoods $\log p(\mathbf{D} \mid \boldsymbol{\theta}, M)$ and prior probabilities $p(\boldsymbol{\theta} \mid M)$ are shown for four MCMC cycles from the power posterior distribution at power coefficient $\beta_1 = 2.13 \times 10^{-8}$. The posterior probability ratio $\alpha_{posterior} = \exp\left[\log(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M) - \log(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M)\right]$, power posterior ratio $\alpha_{power-posterior} = \left(\alpha_{posterior}\right)^{\beta}$, prior probability ratio $\alpha_{prior} = p(\boldsymbol{\theta}_{new} \mid M) / p(\boldsymbol{\theta}_{old} \mid M)$ and the Metropolis acceptance ratio $\alpha = \min\left(1, [\alpha_{power-posterior}\alpha_{prior}]\right)$ are defined in Eq.7. The random number $X \sim U(0,1)$ is generate from uniformly distributed on [0, 1].

	Cycle 1	Cycle 2	Cycle 3	Cycle 4
a_{old}	2.91	2.91	2.91	2.91
m_{old}	2.38	-7.62	-7.62	-7.62
a_{new}	2.91	3.47	4.85	2.82
m_{new}	-7.62	-7.62	-7.62	-6.50
$\log p(\mathbf{D} \mid \boldsymbol{\theta}_{old}, M)$	-1028	-302	-302	-299
$\log p(\mathbf{D} \mid \boldsymbol{\theta}_{new}, M)$	-302	-925	-6291	-250
$\alpha_{posterior}$	Inf	5.03×10^{-271}	0	4.33×10^{-21}
$\alpha_{power-posterior}$	Inf	1.00	0	1.00
$p(\boldsymbol{\theta}_{old} \mid M)$	0.087	3.46×10^{-26}	3.46×10^{-26}	0
$p(\boldsymbol{\theta}_{new} \mid M)$	3.46×10^{-26}	1.77×10^{-26}	9.05×10^{-28}	2.84×10^{-21}
α_{prior}	3.97×10^{-26}	0.51	0.026	2.47×10^{4}
α	1.00	0.51	0	1.00
$X \sim U(0,1)$	0.46	0.73	0.89	0.69
Decision	Accept	Reject	Reject	Accept

These two limits on the Metropolis acceptance ratio have critical implications. While the second threshold merely trims low likelihood regions of the parameter space, more critically the first threshold results in sampling from a distribution that is wider than the target distribution. These two opposing effects result in unwarranted sampling as shown in Figure 6. We show the MCMC sampling of the power posterior distribution at $\beta_1 = 2 \times 10^{-8}$ for the both cases of working with likelihood function and log-likelihood function. We use a sample size of $n = 1 \times 10^{6}$ and burn-in the first 1000 samples. We also plot the contours of the theoretical deviation from the mean. For the case of working with likelihood function (Figure 6c), as we have seen before we cannot sample

all the target distribution due to underflow arising from the lower threshold of $\log(\varepsilon_{dn}) = -745.1332$. However, almost all the generated samples are within the 5σ deviation from the mean, indicating that we are sampling from that target distribution even if we cannot sample all of it. This is not the case for working with log-likelihood as shown in Figure 6d. Similarly, with log-likelihood we cannot sample all the target distribution due to underflow. Yet more critically working with log-likelihood widened the Metropolis acceptance ratio, especially along the direction of parameter m. This resulted in a faster rate of acceptance such that many samples are drawn as from an area as wide as 7σ and 8σ from the mean. These are mainly the samples that have considerable likelihoods, yet low prior probability as previously discussed in Table 4. Accordingly, even though $\alpha_{prior} = p(\boldsymbol{\theta}_{new} \mid M) / p(\boldsymbol{\theta}_{old} \mid M)$ could be small, the upper threshold of $\log(R_{\max}) \approx 709.7827$ will make the Metropolis acceptance ratio overflow to infinity $\alpha = \inf$. Thus, samples as far as 7σ and 8σ from the mean will exist at much higher ratios than their true ratios. This threshold clearly corrupts the Metropolis acceptance ratios and accordingly the MCMC inference. Figure 6d clearly confirms our previous remark that the reduction of TI overestimation when working with the likelihood function is simply a catch. The catch was that the BME overestimation was reduced by some undue low likelihood samples.

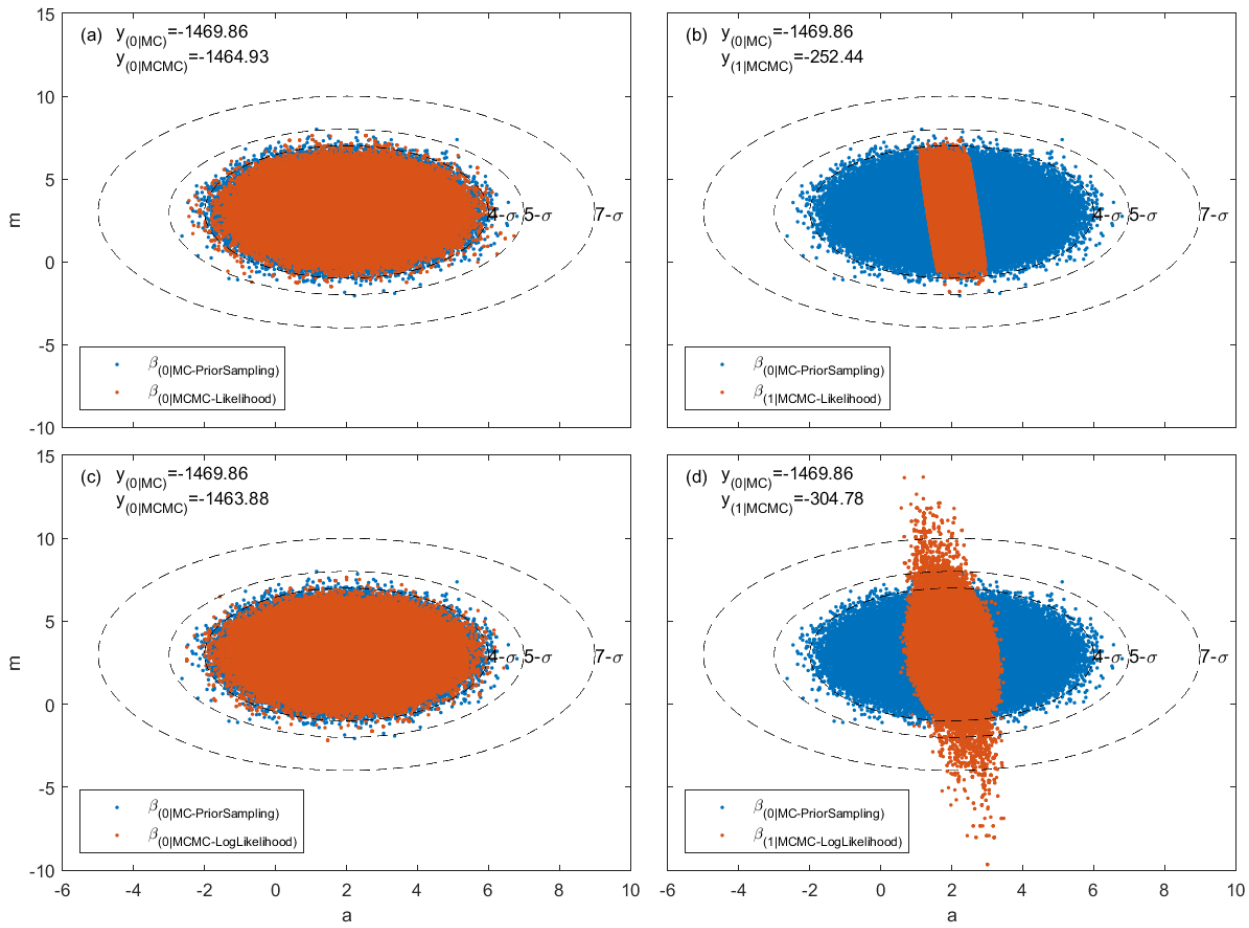

Figure 6. Parameter samples $\boldsymbol{\theta} = \{a, m\}$ with sample size one million draw from the power posterior distributions at (a) $\beta_0 = 0$ (prior) working with likelihood function $\beta_{(0|MCMC-Likelihood)}$, (b) $\beta_0 = 0$ (prior) working with log-likelihood function $\beta_{(0|MCMC-Loglikelihood)}$, (c) $\beta_1 = 2.14 \times 10^{-8}$ working with likelihood function $\beta_{(1|MCMC-Likelihood)}$, (d) $\beta_1 = 2.14 \times 10^{-8}$ working with log-likelihood function $\beta_{(1|MCMC-Loglikelihood)}$. For comparison, we plot MC parameter samples with sample size one million drawn from the prior $\beta_{(0|MC-PriorSampling)}$. For the prior distribution, we also plot the contours of the theoretical deviation from the mean at $4-\sigma$, $5-\sigma$ and $7-\sigma$. We also report the potential expectation y_k (Eq. 13c). Note the reference solutions for y_0 and y_1 are -1469.86 and -1461, respectively.

4.5 Handling of acceptance ratio rounding error by TI and SS

In this discussion we pose the opposite question of why acceptance ratio rounding error that caused TI overestimation did not affect SS. Although TI estimator and SS estimator use the same likelihood values (with the only exception that SS does not need the posterior samples), their

different mathematical formulations determine their vulnerability to acceptance ratio rounding error. For a sequence of $K+1$ distributions, while TI (Eq. 13) requires absolute values of their potential expectation y_k to perform the integration, SS (Eq. 16) merely depends on the multiplication of their importance distribution ratios r_k. This section illustrates how the product of r_k ratios instead of the weighted summation of y_k values is potentially more capable of alleviating the impact of acceptance ratio rounding error up to a certain degree. We show this by focusing on the assignment of β_1 value, as we showed before that β_1 value is a key factor that determines the accuracy of the BME estimation. While SS is relatively invariant to the assignment of β_1 value, TI is very sensitive to the value of β_1. More importantly assignment of β_1 value for TI is a deadlock. For the two cases of $TI_{(0.3,\ 200)}$ and $SS_{(0.3,200)}$, we illustrate two remarks using two problems. These problems are our default informative prior of $a \sim N(2,1)$ and $m \sim N(3,1)$, and a diffuse non-informative prior of $a \sim U(-10,10)$ and $m \sim U(-10,10)$ generated from uniformly distributed on [-10, 10]. For the two cases and two problems, we first generate the β_k values for $K = 200$ and $\alpha = 0.3$ using Eq. 17. Then we estimate the BME by subsequently removing the initial β_k values $\{\beta_1,\ \beta_2\ \ldots.\ \beta_{50}\}$. In other words, for each BME estimation we have $\beta_1^* \in \{\beta_1,\ \beta_2 \ldots \beta_{51}\}$ as a new distribution, where $\beta_1 = 2.14 \times 10^{-8}$ and $\beta_{51} = 0.01$. The results of the informative prior problem (Figure 7a) show that the subsequent removal of the initial β_k values $\{\beta_1,\ \beta_2\ \ldots.\ \beta_{50}\}$ varied the relative error of TI from 92.63% to -99.78%. This is not the case for SS as the relative error varied slightly from -0.03% to 0.89%. For the diffuse non-informative prior problem (Figure 7b) the TI estimation deteriorated greatly, while SS maintained its good performance. Both problems in Figure 7 show clearly that, in the presence of acceptance ratio

rounding error, TI reaches deadlock with respect to the value of β_1. Assigning a very small β_1 value causes BME overestimation (arising from overestimation of y_k values due to acceptance ratio rounding error). Out of the frying pan into the fire, assigning a larger β_1 value results in underestimation as important corners from the integral are missed. That is a double-blind situation in which no decision can be made regarding the value of β_1 (unless we know the true solution).

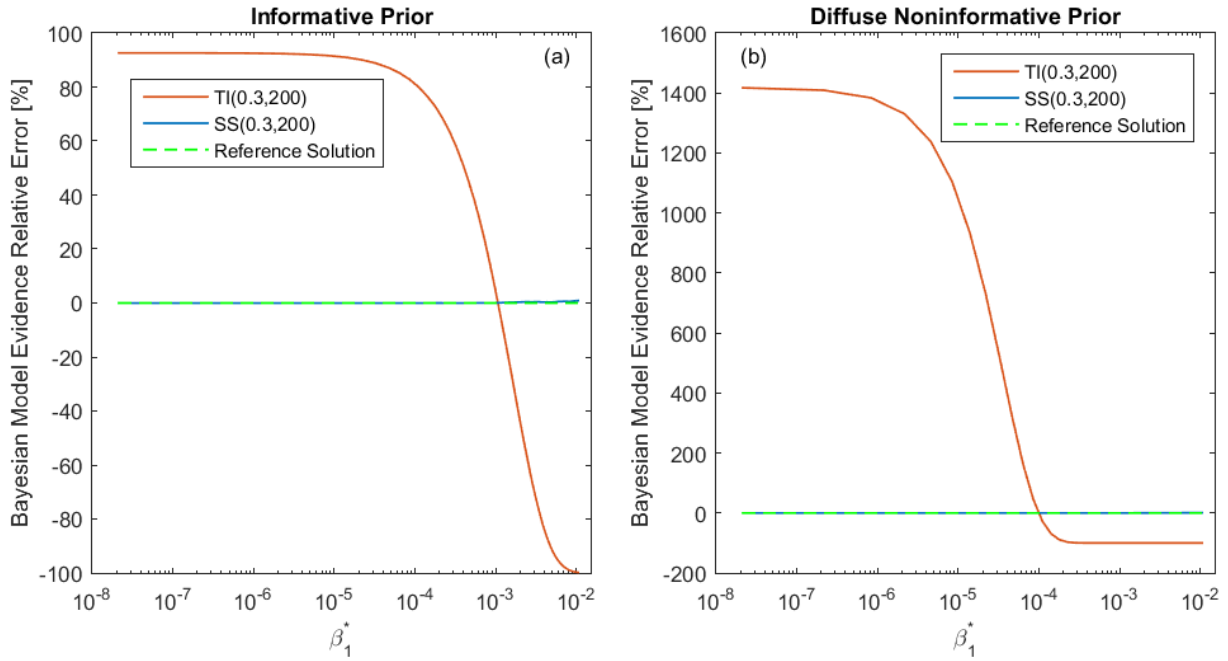

Figure 7. For the linear model with (a) informative prior and (b) diffuse noninformative prior, we show the relative error [%] of a sequence of BME estimations with $\beta_1^* \in \{\beta_1, \beta_2 \ldots \beta_{51}\}$ in which $\beta_1 = 2.14 \times 10^{-8}$ and $\beta_{51} = 0.01$. For both cases of thermodynamic integration $\mathrm{TI}_{(\alpha,K)}$ and steppingstone sampling $SS_{(\alpha,K)}$, the initial power posterior coefficients $\beta_k \in \{\beta_0, \beta_1 \ldots \beta_K\}$ are generated using Eq.17 with shape parameter $\alpha = 0.3$ and $K = 200$. Note the BME reference solutions for the two problems of informative prior and diffuse noninformative prior are 5.1194×10^{-14} and 8.8757×10^{-16}, respectively. The reference solution of the relative error [%] is zero by definition.

One remaining point is to show why SS did not fail. Given $SS_{(0.3,200)}$, Figure 8 plots the impotence sampling ratios $r_{ss,k}$ with β_k values. We show this for the default case of using all the β values $SS_{(0.3,200)} \forall \beta$. We also show the case of removing the initial β_k values that are subject to

acceptance ratio rounding error $SS^{-}_{(0.3,200)} \forall \beta - \{\beta_1 = 2.14 \times 10^{-8}, ..., 9.77 \times 10^{-4} = \beta_{25}\}$ such that $\beta_{26} = 1.11 \times 10^{-4}$ becomes the new β_1 instead of $\beta_1 = 2.14 \times 10^{-8}$. Interestingly, the two curves overlap perfectly in the two cases indicating that these problematic ratios have no impact. We further illustrate this by multiplying the $r_{ss,k}$ ratios up to $\beta = 1.11 \times 10^{-4}$. Interestingly, the multiplication result $r_{SS} = \prod_{k=1}^{26} r_{ss,k} = 0.3993$ of the default case $SS_{(0.3,200)}$, and the multiplication result $r_{SS} = \prod_{k=1}^{1} r_{ss,k} = 0.4083$ of the special case $SS^{-}_{(0.3,200)}$ tuned out to be almost identical. It seems that the systematic bias introduced by acceptance ratio rounding error is factored out through the multiplication of $r_{ss,k}$ ratios, but will definitely adds up in the weighted summation of y_k values.

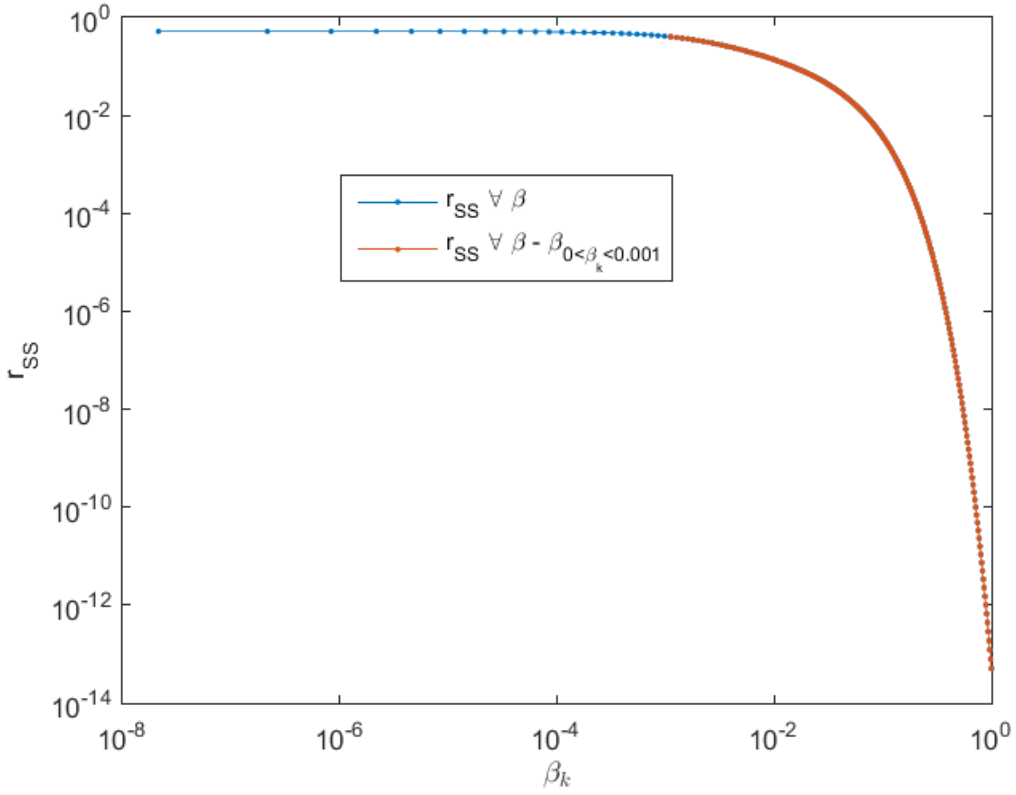

Figure 8. For the linear model, we show the impotence sampling ratios $r_{ss,k}$ of steppingstone sampling (SS) at different β_k values, which are generated using Eq.17 with shape parameter $\alpha = 0.3$ and $K = 200$. We report $r_{ss,k}$ ratios for the default case of using all the β values $SS_{(0.3,200)} \forall \beta$, and for the special case of removing the initial β_k values that are subject to rounding sampling error $SS^{-}_{(0.3,200)} \forall \beta - \{\beta_1 = 2.14 \times 10^{-8}, ..., 9.77 \times 10^{-4} = \beta_{25}\}$.

5. Discussion and conclusions

Estimating the Bayesian model evidence (BME) is a challenging task in Bayesian computation. To ensure reliable model selection, BME numerical estimators are gaining more attention in various science areas due to increased accuracy over semi-analytical solutions such as Laplace approximations. Theoretically, numerical estimators are bias-free, as they holds no underlying assumptions about the shapes of the prior, likelihood or posterior distributions. However, in most practical applications these distributions need to be approximated using MCMC for methods such as harmonic mean (HM), thermodynamic integration (TI), and steppingstone

sampling (SS), or with MC simulation for the arithmetic mean (AM) estimator. These numerical approximations can introduce several bias sources, and more knowledge about errors of numerical estimators is needed to better understand and reduce the uncertainty of model selection and model averaging when using numerical estimators. In this work we outline potential error sources of numerical estimators. Specifically, we analyze in detail the likelihood rounding error and the acceptance ratio rounding error, which have not previously discussed in literature. We show that the likelihood rounding error can result in severe underestimation of BME, while the acceptance ratio rounding error can result in overestimation of BME. Thus, these two error sources can introduce bias that potentially changes the model ranking. Acceptance ratio rounding error could occur during the MCMC sampling process, and likelihood rounding error could occur in the post-analysis when using likelihoods with different estimators to estimate the BME. These two numerical daemons arise from arithmetic precision of floating point numbers. This manuscript illustrates the conditions that invoke these daemons and their implications on BME numerical estimation.

The study shows that rounding error can be avoided by a theoretical development that leads to advance mathematical formulations. AM estimator is indeed the simplest numerical approximation of the BME integral (Eq.4), yet the most susceptible to underflow of the likelihoods due to the likelihood rounding error. More advanced numerical approximations of the BME integral such as TI and SS are less susceptible to likelihood rounding error. Unfortunately, most practical applications require MCMC sampling, and accordingly the acceptance ratio rounding error associated with the Metropolis acceptance ratios acceptance ratio rounding imposes a limit on the benefits gained by the elaborate mathematical formulations of TI and SS.

The study provides two main insights that need to be considered for designing unbiased BME numerical estimators. First, likelihood rounding error does not appear when we post-process the log-likelihoods to estimate BME (e.g., TI), but can potentially appear when we post-process the likelihoods to estimate BME (e.g., AH, HM and SS). The prior sampling AM estimator is the most susceptible to likelihood rounding error due to subnormal underflow of likelihoods. This is extremely unlikely to occur with SS estimator as the likelihoods are raised to small fraction. This is also extremely unlikely to occur with the posterior sampling HM estimator as posterior samples would generally have likelihoods that are much higher than subnormal underflow threshold. Second, acceptance ratio rounding error can potentially occur during MCMC sampling introducing a numerical threshold when working with likelihood function and two thresholds when working with log-likelihood function. For working with likelihood function, this numerical threshold can trim the low likelihood regions of the parameter space. Working with log-likelihood while considering a non-uniform prior, can additionally corrupt the empirical sampling ratios of the MCMC sampler.

With respect practical implementation of numerical methods, this study provides two suggestions. First, these numerical artifacts associated with log-likelihood function can occur for distributions that are slightly less dispersed than the prior, and are unlikely to occur when estimating the posterior or for distributions that are slightly more dispersed than the posterior. Thus, we conclude that, while it is generally recommend to work with log-likelihood for Bayesian inference through posterior sampling, it is not recommend to work with log-likelihood for BME estimation without checking for rounding error. Second, checking for rounding error is straightforward. We can calculate the percentage of samples that underflow to zero. This can be reported for a more transparent BME estimation process. We can additionally plot parameter

distributions of different power posterior distributions (e.g., Figure 6) as a regular check for gradual unimpaired sampling from the prior to the posterior.

With respect to these two rounding error artifacts, the study also shows several estimator specific findings. With respect to AM, we refute the assumption that AM estimator is bias-free estimator that will approximate the true solution by investing in computational effort. We show that, in the presence of likelihood rounding error, the AM estimator will converge to a smallest nonzero denormalized floating-point number, but not to the true solution. Nevertheless, it is still valid to use AM as the reference solution in BME benchmarking studies, provided that a check for likelihood rounding error is carried out as AM can be a highly biased estimator. With respect to HM, unlike prior sampling AM estimator the posterior sampling HM estimator is the least vulnerable to numerical demons as it evades low likelihood regions in the parameter space. For these two different reasons, the AM underestimation bias and the HM overestimation bias increase for high dimensional problems with a peaked likelihood functions. This predicament between prior and posterior sampling impacts TI in a different way. Sampling likelihood error creates an interesting deadlock for the TI estimator such that breaking this deadlock is like choosing between a rock and a hard place. In other words, avoiding distributions that are slightly less dispersed than the prior results in BME underestimation by cutting corners from the integral, while by considering them we risk encountering acceptance ratio rounding error that results in BME overestimation. Finally, by taking ratios of importance distributions rather than integrating expected likelihoods using TI, the SS can potentially be less vulnerable to these numerical challenges.

In conclusion, by knowing the impacts of numerical daemons of BME numerical estimators, we can anticipate for their adverse effects. This study generally enrich the MCMC error analysis by bring attention to other error sources that are liable to occur when sampling any

dispersed target distribution with low likelihood regions. We hope that this contribution shall enlighten the developers and users of BME numerical estimators that numerical daemons are lurking, and provide them with insights to spot for theses daemons.

Acknowledgement

This research was supported by NSF grant EAR-1552329.

Code availability

Data that support the findings of this study are publicly available (Elshall 2021). The codes that are used to produce this paper are available from the corresponding author at mye@fsu.edu. We cannot publicly share the workflow because the MT-DREAM(ZS) code (Laloy and Vrugt 2012), which is a main component in the workflow, is in the process of becoming a commercial code.

Authors contribution

ASE: Conceptualization, Methodology, Data curation, Software, Investigation, Visualization, Validation, Writing - original draft, Writing - review & editing. MY: Conceptualization, Methodology, Investigation, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

References

Baele, G., Lemey, P., Vansteelandt, S.: Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. BMC Bioinformatics. 14, 85 (2013). https://doi.org/10.1186/1471-2105-14-85

Calderhead, B., Girolami, M.: Estimating Bayes factors via thermodynamic integration and population MCMC. Computational Statistics & Data Analysis. 53, 4028–4045 (2009). https://doi.org/10.1016/j.csda.2009.07.025

Chitsazan, N., Nadiri, A.A., Tsai, F.T.C.: Prediction and structural uncertainty analyses of artificial neural networks using hierarchical Bayesian model averaging. Journal of Hydrology. 528, 52–62 (2015). https://doi.org/10.1016/j.jhydrol.2015.06.007

Clark, M.P., Kavetski, D.: Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. Water Resources Research. 46, (2010). https://doi.org/10.1029/2009WR008894

Elshall, A.S.: Data of numerical daemons in Monte Carlo estimators of Bayesian model evidence. (2021). https://doi.org/10.5281/zenodo.5121596

Elshall, A.S., Ye, M., Pei, Y., Zhang, F., Niu, G.-Y.G.-Y., Barron-Gafford, G.A.: Relative model score: a scoring rule for evaluating ensemble simulations with application to microbial soil respiration modeling. Stochastic Environmental Research and Risk Assessment. 32, 2809–2819 (2018). https://doi.org/10.1007/s00477-018-1592-3

Elshall, Ye: Making Steppingstones out of Stumbling Blocks: A Bayesian Model Evidence Estimator with Application to Groundwater Transport Model Selection. Water. 11, 1579 (2019). https://doi.org/10.3390/w11081579

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., Lewis, P.O.: Choosing among Partition Models in Bayesian Phylogenetics. Mol Biol Evol. 28, 523–532 (2011). https://doi.org/10.1093/molbev/msq224

Friel, N., Hurn, M., Wyse, J.: Improving power posterior estimation of statistical evidence. Stat Comput. 24, 709–723 (2014). https://doi.org/10.1007/s11222-013-9397-1

Friel, N., Pettitt, A.N.: Marginal likelihood estimation via power posteriors. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 70, 589–607 (2008). https://doi.org/10.1111/j.1467-9868.2007.00650.x

Friel, N., Wyse, J.: Estimating the evidence – a review. Statistica Neerlandica. 66, 288–308 (2012). https://doi.org/10.1111/j.1467-9574.2011.00515.x

Gelman, A., Meng, X.-L.: Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Statistical Science. 13, 163–185 (1998). https://doi.org/10.1214/ss/1028905934

Heath, M.T.: Scientific computing: an introductory survey. McGraw-Hill, Boston (1997)

Kavetski, D., Clark, M.P.: Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. Water Resources Research. 46, (2010). https://doi.org/10.1029/2009WR008896

La Follette, P.T., Teuling, A.J., Addor, N., Clark, M., Jansen, K., Melsen, L.A.: Numerical daemons of hydrological models are summoned by extreme precipitation. Hydrology and Earth System Sciences Discussions. 1–32 (2021). https://doi.org/10.5194/hess-2021-28

Laloy, E., Vrugt, J.A.: High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing. Water Resources Research. 48, (2012). https://doi.org/10.1029/2011WR010608

Lartillot, N., Philippe, H.: Computing Bayes Factors Using Thermodynamic Integration. Systematic Biology. 55, 195–207 (2006). https://doi.org/10.1080/10635150500433722

Lefebvre, G., Steele, R., Vandal, A.C.: A path sampling identity for computing the Kullback–Leibler and J divergences. Computational Statistics & Data Analysis. 54, 1719–1731 (2010). https://doi.org/10.1016/j.csda.2010.01.018

Liu, P., Elshall, A.S.A.S., Ye, M., Beerli, P., Zeng, X., Lu, D., Tao, Y.: Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. Water Resources Research. 52, 734–758 (2016). https://doi.org/10.1002/2014WR016718

Lu, D., Ye, M., Curtis, G.P.: Maximum likelihood Bayesian model averaging and its predictive analysis for groundwater reactive transport models. Journal of Hydrology. 529, 1859–1873 (2015). https://doi.org/10.1016/j.jhydrol.2015.07.029

Lu, D., Ye, M., Hill, M.C.: Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification. Water Resources Research. 48, (2012). https://doi.org/10.1029/2011WR011289

Marshall, L., Nott, D., Sharma, A.: Hydrological model selection: A Bayesian alternative. Water Resources Research. 41, (2005). https://doi.org/10.1029/2004WR003719
McCulloch, R., Rossi, P.E.: A bayesian approach to testing the arbitrage pricing theory. Journal of Econometrics. 49, 141–168 (1991). https://doi.org/10.1016/0304-4076(91)90012-3
Newton, M.A., Raftery, A.E.: Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society. Series B (Methodological). 56, 3–48 (1994)
Pham, H.V., Tsai, F.T.-C.: Optimal observation network design for conceptual model discrimination and uncertainty reduction: Observtion network design for model discrimination. Water Resources Research. 52, 1245–1264 (2016). https://doi.org/10.1002/2015WR017474
Recktenwald, G.W.: Numerical Methods with MATLAB: Implementations and Applications. Prentice Hall, New Jersey (2000)
Rojas, R., Kahunde, S., Peeters, L., Batelaan, O., Feyen, L., Dassargues, A.: Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling. Journal of Hydrology. 394, 416–435 (2010). https://doi.org/10.1016/j.jhydrol.2010.09.016
Schöniger, A., Wöhling, T., Nowak, W.: A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. Water Resources Research. 51, 7524–7546 (2015). https://doi.org/10.1002/2015WR016918
Schöniger, A., Wöhling, T., Samaniego, L., Nowak, W.: Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resources Research. 50, 9484–9513 (2014). https://doi.org/10.1002/2014WR016062
Vrugt, J.A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling & Software. 75, 273–316 (2016). https://doi.org/10.1016/j.envsoft.2015.08.013
Wöhling, T., Schöniger, A., Gayler, S., Nowak, W.: Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. Water Resources Research. 51, 2825–2846 (2015). https://doi.org/10.1002/2014WR016292
Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.-H.: Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst Biol. 60, 150–160 (2011). https://doi.org/10.1093/sysbio/syq085
Zhang, X., Niu, G.-Y., Elshall, A.S., Ye, M., Barron-Gafford, G.A., Pavao-Zuckerman, M.: Assessing five evolving microbial enzyme models against field measurements from a semiarid savannah—What are the mechanisms of soil respiration pulses? Geophysical Research Letters. 41, 6428–6434 (2014). https://doi.org/10.1002/2014GL061399