Properties of nested sampling

Author(s): NICOLAS CHOPIN and CHRISTIAN P. ROBERT

Source: *Biometrika*, SEPTEMBER 2010, Vol. 97, No. 3 (SEPTEMBER 2010), pp. 741-755

Published by: Oxford University Press on behalf of Biometrika Trust

Stable URL: https://www.jstor.org/stable/25734120

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/25734120?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

and *Oxford University Press* are collaborating with JSTOR to digitize, preserve and extend access to *Biometrika*

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

Biometrika (2010), **97**, 3, *pp.* 741–755
doi: 10.1093/biomet/asq021
Advance Access publication 1 June 2010
© 2010 Biometrika Trust
Printed in Great Britain

Properties of nested sampling

BY NICOLAS CHOPIN

CREST–ENSAE, Timbre J120, 3, Avenue Pierre Larousse, 92245 Malakoff cedex, France
nicolas.chopin@ensae.fr

AND CHRISTIAN P. ROBERT

Université Paris-Dauphine, CEREMADE, F-775 Paris cedex 16, France
xian@ceremade.dauphine.fr

SUMMARY

Nested sampling is a simulation method for approximating marginal likelihoods. We establish that nested sampling has an approximation error that vanishes at the standard Monte Carlo rate and that this error is asymptotically Gaussian. It is shown that the asymptotic variance of the nested sampling approximation typically grows linearly with the dimension of the parameter. We discuss the applicability and efficiency of nested sampling in realistic problems, and compare it with two current methods for computing marginal likelihood. Finally, we propose an extension that avoids resorting to Markov chain Monte Carlo simulation to obtain the simulated points.

Some key words: Central limit theorem; Evidence; Importance sampling; Marginal likelihood; Markov chain Monte Carlo simulation; Nested sampling.

1. INTRODUCTION

Nested sampling was introduced by Skilling (2006) as a numerical approximation method for integrals:

$$Z = \int L(y \mid \theta)\pi(\theta)\,\mathrm{d}\theta,$$

when π is the prior distribution and $L(y \mid \theta)$ is the likelihood. Those integrals are called evidence in the above paper. They naturally occur as marginals in Bayesian testing and model choice (Jeffreys, 1939; Robert, 2001, Chs. 5 and 7). Nested sampling has been well received in astronomy and has been applied successfully to several cosmological problems; see, Mukherjee et al. (2006), Shaw et al. (2007) and Vegetti & Koopmans (2009), among others. In addition, Murray et al. (2006) developed a nested sampling algorithm for computing the normalizing constant of Potts models.

The purpose of this paper is to investigate the formal properties of nested sampling. Evans (2007) showed that nested sampling estimates converge in probability, but he calls for further work on the rate of convergence and the limiting distribution. Our main result is a central limit theorem for nested sampling estimates, stating that the approximation error is dominated by an $O(N^{-1/2})$ stochastic term, with a limiting Gaussian distribution, where N is a tuning parameter proportional to the computational effort. We also investigate the impact of the dimension d of the problem on the performances of the algorithm. In a simple example, we show that the asymptotic variance of nested sampling estimates grows linearly with d; this means that the computational cost is $O(d^3/\eta^2)$, where η is the selected error bound.

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

One important aspect of nested sampling is that it resorts to simulating points θ_i from the prior π, constrained to θ_i having a larger likelihood value than some threshold l. In many cases, the simulated points must be generated by Markov chain Monte Carlo sampling. We propose an extension of nested sampling, based on importance sampling as in Mukherjee et al. (2006), that introduces enough flexibility to perform the constrained simulation without resorting to Markov chain Monte Carlo methods.

Finally, we examine two alternatives to nested sampling, both based on the output of Markov chain Monte Carlo algorithms. We do not aim at an exhaustive comparison with all existing methods; see, for instance, Chen et al. (2000) for a broader review, and restrict the comparison to methods that share with nested sampling the property that they provide approximations of both the posterior distribution and the marginal likelihood, at no extra cost. We provide numerical comparisons between these methods. A 2007 PhD thesis by Murray at University College London also includes numerical comparisons of nested sampling with other methods for several models.

2. NESTED SAMPLING: A DESCRIPTION

2·1. *Principle*

We briefly describe the nested sampling algorithm, as introduced by Skilling (2006). We use $L(\theta)$ as a shorthand for the likelihood $L(y \mid \theta)$, omitting the dependence on y.

Nested sampling is based on the identity

$$Z = \int_0^1 \varphi(x)\,\mathrm{d}x,$$

where φ is the tail quantile function of the random variable $L(\theta)$, associated with its survival function, $\varphi^{-1}(l) = \mathrm{pr}\{L(\theta) > l\}$, assuming $\theta \sim \pi$ and φ is a decreasing function, which is the case when L is continuous and π has a connected support. The representation $Z = E^{\pi}\{L(\theta)\}$ holds with no restriction on either L or π. Formally, this one-dimensional integral could be approximated by standard quadrature methods,

$$\hat{Z} = \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i, \tag{1}$$

where $\varphi_i = \varphi(x_i)$, and $0 < x_j < \cdots < x_1 < x_0 = 1$ is an arbitrary grid over $[0, 1]$. The function φ is intractable in most cases however, so the φ_is are approximated by an iterative random mechanism as follows.

Step 1. Draw independently N points $\theta_{1,i}$ from the prior distribution π, determine $\theta_1 = \arg\min_{i=1,\ldots,N} L(\theta_{1,i})$, and set $\varphi_1 = L(\theta_1)$.

Step 2. Obtain the N current values $\theta_{2,i}$, by reproducing the $\theta_{1,i}$s, except for θ_1 which is replaced by a draw from the prior distribution π conditional upon $L(\theta) \geqslant \varphi_1$; then select θ_2 as $\theta_2 = \arg\min_{i=1,\ldots,N} L(\theta_{2,i})$, and set $\varphi_2 = L(\theta_2)$.

Step 3. Repeat Steps 1 and 2 until a given stopping rule is satisfied, for instance until observing very small changes in the approximation $\hat{Z}$ or until reaching the maximal value of $L(\theta)$ when it is known.

In the above, the values $x_i^{\star} = \varphi^{-1}(\varphi_i)$ that appear in the quadrature approximation (1) are unknown, but they have the following property: $t_i = \varphi^{-1}(\varphi_{i+1})/\varphi^{-1}(\varphi_i) = x_{i+1}^{\star}/x_i^{\star}$ are independent $\mathrm{Be}(N, 1)$ variates. Skilling (2006) proposes two approaches: first, a deterministic scheme, where

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

x_i is replaced with $\exp(-i/N)$ in (1), so that $\log x_i$ is the expectation of $\log \varphi^{-1}(\varphi_i)$; second, a random scheme, where K parallel streams of random numbers $x_{i,k}$ $(k = 1, \ldots, K)$ are generated from the same generating process as the $x_i^\star$s, $x_{i+1,k} = x_{i,k} t_{i,k}$, where $t_{i,k} \sim \mathrm{Be}(N, 1)$. In the latter case, a natural estimator is

$$\log \tilde{Z} = \frac{1}{K} \sum_{k=1}^{K} \log \tilde{Z}_k, \quad \tilde{Z}_k = \sum_{i=1}^{j} (x_{i-1,k} - x_{i,k}) \varphi_i.$$

For the sake of brevity, we focus on the deterministic scheme in this paper, and study the estimator (1) and $x_i = \exp(-i/N)$. For $K = 1$, the random scheme produces more noisy estimates than the deterministic scheme, but, for large values of K, the opposite may occur; see, for instance, Fig. 3 in Murray et al. (2006).

2·2. *Variations and posterior simulation*

Skilling (2006) indicates that nested sampling provides simulations from the posterior distribution at no extra cost: "the existing sequence of points $\theta_1, \theta_2, \theta_3, \ldots$ already gives a set of posterior representatives, provided the ith is assigned the appropriate importance weight $\omega_i L_i$", where the weight ω_i is equal to the difference $(x_{i-1} - x_i)$ and L_i is equal to φ_i. This can be justified as follows. Consider the computation of the posterior expectation of a given function f:

$$\mu(f) = \int \pi(\theta) L(\theta) f(\theta) \, \mathrm{d}\theta \Big/ \int \pi(\theta) L(\theta) \, \mathrm{d}\theta.$$

One can then use a single run of nested sampling to obtain estimates of both the numerator and the denominator, the latter being the evidence Z, estimated by (1). The estimator

$$\sum_{i=1}^{j} (x_{i-1} - x_i) \varphi_i f(\theta_i) \tag{2}$$

of the numerator is a noisy version of $\sum_{i=1}^{j} (x_{i-1} - x_i) \varphi_i \tilde{f}(\varphi_i)$, where $\tilde{f}(l) = E^\pi \{ f(\theta) \mid L(\theta) = l \}$, the prior expectation of $f(\theta)$ conditional on $L(\theta) = l$. This Riemann sum is, following the principle of nested sampling, an estimator of the evidence.

LEMMA 1. *Let $\tilde{f}(l) = E^\pi \{ f(\theta) \mid L(\theta) = l \}$ for $l > 0$. Then, if $\tilde{f}$ is absolutely continuous,*

$$\int_0^1 \varphi(x) \tilde{f}\{\varphi(x)\} \, \mathrm{d}x = \int \pi(\theta) L(\theta) f(\theta) \, \mathrm{d}\theta. \tag{3}$$

A proof is provided in the Appendix. Clearly, the estimate of $\mu(f)$ obtained by dividing (2) by (1) is the estimate obtained by computing the weighted average mentioned above. We do not discuss further this aspect of nested sampling, but our convergence results can be extended to such estimates.

3. A CENTRAL LIMIT THEOREM FOR NESTED SAMPLING

We decompose the approximation error of nested sampling as follows:

$$\sum_{i=1}^{j} (x_{i-1} - x_i) \varphi_i - \int_0^1 \varphi(x) \, \mathrm{d}x = -\int_0^\varepsilon \varphi(x) \, \mathrm{d}x + \left\{ \sum_{i=1}^{j} (x_{i-1} - x_i) \varphi(x_i) - \int_\varepsilon^1 \varphi(x) \, \mathrm{d}x \right\}$$
$$+ \sum_{i=1}^{j} (x_{i-1} - x_i) \{ \varphi_i - \varphi(x_i) \}.$$

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

The first term is a truncation error, resulting from the feature that the algorithm is run for a finite time. For simplicity's sake, we assume that the algorithm is stopped at iteration $j = \lceil(-\log\varepsilon)N\rceil$, where $\lceil x\rceil$ stands for the smallest integer k such that $x \leqslant k$, so that $x_j = \exp(-j/N) \leqslant \varepsilon < x_{j-1}$. More practical stopping rules are discussed in § 7. Assuming φ, or equivalently L, bounded from above, the error $\int_0^\varepsilon \varphi(x)\,\mathrm{d}x$ is exponentially small with respect to the computational effort. The second term is a numerical integration error, which, provided φ' is bounded over $[\varepsilon, 1]$, is of order $O(N^{-1})$, since $x_{i-1} - x_i = O(N^{-1})$. The third term is stochastic and is denoted by

$$\eta_N = \sum_{i=1}^{j}(x_{i-1} - x_i)\{\varphi(x_i^\star) - \varphi(x_i)\},$$

where the $x_i^\star$s are such that $\varphi_i = L(\theta_i) = \varphi(x_i^\star)$, therefore $x_i^\star = \varphi^{-1}(\varphi_i)$.

The following theorem characterizes the asymptotic behaviour of η_N.

THEOREM 1. *If φ is twice continuously differentiable over $[\varepsilon, 1]$, and if its two first derivatives are bounded over $[\varepsilon, 1]$, then $N^{1/2}\eta_N$ converges in distribution to a Gaussian distribution with mean zero and variance*

$$V = -\int_{s,t\in[\varepsilon,1]} s\varphi'(s)t\varphi'(t)\log(s \vee t)\,\mathrm{d}s\,\mathrm{d}t.$$

The stochastic error is of order $O_P(N^{-1/2})$ and it dominates both the other error terms. The proof of this theorem relies on the functional central limit theorem and is detailed in the Appendix. A straightforward application of the delta method shows that the log-scale error, $\log\hat{Z} - \log Z$, has the same asymptotic behaviour, but with asymptotic variance V/Z^2.

4. PROPERTIES OF THE NESTED SAMPLING ALGORITHM

4·1. *Simulating from a constrained prior*

The main difficulty of nested sampling is to simulate θ from the prior distribution π subject to the constraint $L(\theta) > L(\theta_i)$; exact simulation from this distribution is intractable in many realistic set-ups. Nested sampling is at least of the same complexity as a one-dimensional slice sampler, which produces a uniformly ergodic Markov chain when the likelihood L is bounded but may be slow to converge in other settings (Roberts & Rosenthal, 1999).

Skilling (2006) proposes sampling values of θ by iterating M Markov chain Monte Carlo steps, using the truncated prior as the invariant distribution, and a point chosen at random among the $N - 1$ survivors as the starting point. Since the starting value is distributed from the invariant distribution, a finite number M of iterations produces an outcome that is marginally distributed from the correct distribution. This, however, introduces correlations between simulated points. We stress that our central limit theorem does not apply when the simulated points are not independent, and that the consistency of nested sampling estimates based on Markov chain Monte Carlo is an open problem. A theoretical result seems difficult to establish because each iteration involves both a different Markov chain Monte Carlo kernel and a different invariant distribution.

There are settings when implementing a Markov chain Monte Carlo move that leaves the truncated prior invariant is convoluted. In such cases, one may instead implement a Markovian transition, for instance a random walk Metropolis–Hastings move, with respect to the unconstrained prior, and subsample only values that satisfy the constraint $L(\theta) > L(\theta_i)$, but this scheme may become increasingly inefficient as the constraint moves closer to the highest values of L. More

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

advanced sampling schemes can be devised that overcome this difficulty, such as the use of a diminishing variance factor in the random walk.

In § 5, an extension of nested sampling is derived from importance sampling. In some settings, this may facilitate the design of efficient Markovian moves, or even allow for sampling independently the θ_is.

4·2. *The impact of dimensionality*

We show in this section that the theoretical performance of nested sampling typically depends on the dimension d of the problem as follows: the required number of iterations and the asymptotic variance both grow linearly with d. Thus, if a single iteration costs $O(d)$, the computational cost of nested sampling is $O(d^3/\eta^2)$, where η denotes a given error level. Murray's PhD thesis also states this result, using a more heuristic argument. This result applies to the exact nested algorithm only. In principle, resorting to Markov chain Monte Carlo methods might entail some additional curse of dimensionality, but this point seems difficult to study formally, and will only be briefly investigated in our simulation studies.

Consider the case where, for $k = 1, \ldots, d$, $\theta^{(k)} \sim \mathcal{N}(0, \sigma_0^2)$ and $y^{(k)} \mid \theta^{(k)} \sim \mathcal{N}(\theta^{(k)}, \sigma_1^2)$, independently in both cases. Set $y^{(k)} = 0$ and $\sigma_0^2 = \sigma_1^2 = 1/4\pi$, so that $Z = 1$ for all ds. A draw from the constrained prior is obtained as follows: simulate $r^2 \leqslant -2\log\{l2^{-d/2}\}$ from a truncated $\chi^2(d)$ distribution and $u_1, \ldots, u_d \sim \mathcal{N}(0, 1)$, then set $\theta^{(k)} = \sigma_0 r\, u_k/(u_1^2 + \cdots + u_d^2)^{1/2}$. Since $Z = 1$, we assume that the truncation point ε_d is such that $\varphi(0)\varepsilon_d = \tau \ll 1$, $\tau = 10^{-6}$, say, where $\varphi(0) = 2^{d/2}$ is the maximum likelihood value. Therefore, $\varepsilon_d = \tau 2^{-d/2}$ and the number of iterations required to produce a given truncation error, that is, $j = \lceil(-\log\epsilon)N\rceil$, grows linearly in d. To assess the dependence of the asymptotic variance with respect to d, we state the following lemma, established in the Appendix.

LEMMA 2. *In the current setting, if V_d is the asymptotic variance of the nested sampling estimator with truncation point ε_d, there exist constants c_1, c_2 such that $V_d/d \leqslant c_1$ for all $d \geqslant 1$, and* $\liminf_{d\to+\infty} V_d/d \geqslant c_2$.

This lemma is easily generalized to cases where the prior is such that the components are independent and identically distributed, and the likelihood factorizes as $L(\theta) = \prod_{k=1}^{d} L(\theta^{(k)})$. We conjecture that V_d/d converges to a finite value in all these situations and that, for more general models, V_d grows linearly with the dimensionality of the problem, as measured, for instance, in Spiegelhalter et al. (2002).

5. NESTED IMPORTANCE SAMPLING

We introduce an extension of nested sampling based on importance sampling. Let $\tilde{\pi}(\theta)$ denote an instrumental prior with the support of π included in the support of $\tilde{\pi}$, and let $\tilde{L}(\theta)$ denote an instrumental likelihood, namely a positive measurable function. We define an importance weight function $w(\theta)$ such that $\tilde{\pi}(\theta)\tilde{L}(\theta)w(\theta) = \pi(\theta)L(\theta)$. We can approximate Z by nested sampling for the pair $(\tilde{\pi}, \tilde{L})$, that is, by simulating iteratively from $\tilde{\pi}$ constrained to $\tilde{L}(\theta) > l$, and by computing the generalized nested sampling estimator

$$\sum_{i=1}^{j}(x_{i-1} - x_i)\varphi_i w(\theta_i).$$

The advantage of this extension is that one can choose $(\tilde{\pi}, \tilde{L})$ so that simulating from $\tilde{\pi}$ under the constraint $\tilde{L}(\theta) > l$ is easier than simulating from π under the constraint $L(\theta) > l$. For instance,

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

one may choose an instrumental prior $\tilde{\pi}$ such that Markov chain Monte Carlo steps adapted to the instrumental constrained prior are easier to implement than with respect to the actual constrained prior. In a similar vein, nested importance sampling facilitates contemplating several priors at once, as one may compute the evidence for each prior by producing the same nested sequence, based on the same pair $(\tilde{\pi}, \tilde{L})$, and by simply modifying the weight function.

Ultimately, one may choose $(\tilde{\pi}, \tilde{L})$ so that the constrained simulation is performed exactly. For instance, if $\tilde{\pi}$ is a Gaussian $\mathcal{N}_d(\hat{\theta}, \hat{\Sigma})$ distribution with arbitrary hyperparameters, take

$$\tilde{L}(\theta) = \lambda\{(\theta - \hat{\theta})^{\mathrm{T}} \hat{\Sigma}^{-1} (\theta - \hat{\theta})\},$$

where λ is an arbitrary decreasing function. Then

$$\varphi_i w(\theta_i) = \tilde{L}(\theta_i) w(\theta_i) = \pi(\theta_i) L(\theta_i) / \tilde{\pi}(\theta_i)$$

and the x_is in (1) are error-free: at iteration i, θ_i is sampled uniformly over the ellipsoid with an $\exp(-i/N)$ prior mass as $\theta_i = \hat{\theta} + q_i^{1/2} C v / \|v\|_2^{1/2}$, where C is the Cholesky lower triangle of $\hat{\Sigma}$, $v \sim N_d(0, I_d)$ and q_i is the $\exp(-i/N)$ quantile of a $\chi^2(d)$ distribution. Mukherjee et al. (2006) consider a nested sampling algorithm where simulations are generated within an ellipsoid, and accepted if they fit the likelihood constraint, but their algorithm differs from the importance sampling extension described here.

The nested ellipsoid strategy seems useful in two scenarios. First, assume that both the posterior mode and the Hessian at the mode are available numerically and tune $\hat{\theta}$ and $\hat{\Sigma}$ accordingly. In this case, this strategy should outperform standard importance sampling based on the optimal Gaussian proposal, because the nested ellipsoid strategy uses an $O(N^{-1})$ quadrature rule on the radial axis, along which the weight function varies the most; see § 7·3 for an illustration. Second, assume that only the posterior mode is available, so one may set $\hat{\theta}$ to the posterior mode, and set $\hat{\Sigma} = \tau I_d$, where τ is an arbitrary, large value. Section 7·3 indicates that the nested ellipsoid strategy may still perform reasonably in such a scenario. Models for which the Hessian at the mode is tedious to compute include in particular Gaussian state space models with missing observations (Frühwirth-Schnatter, 2004), Markov modulated Poisson processes (Rydén, 1994), or, more generally, models where the expectation-maximization algorithm (MacLachlan & Krishnan, 1997) is the easiest way to compute the posterior mode, although one may use Louis' (1982) method for computing the information matrix from the expectation-maximization output.

6. ALTERNATIVE ALGORITHMS

6·1. *Approximating Z from a posterior sample*

As recalled in § 2·2, the output of nested sampling can be recycled to approximate posterior quantities. Conversely, one can recycle the output of a Markov chain Monte Carlo algorithm to estimate the evidence, with no or little additional programming effort; see, for instance, Gelfand & Dey (1994), Meng & Wong (1996) and Chen & Shao (1997). We describe below the solutions used in the subsequent comparison with nested sampling, but we do not give an exhaustive coverage of those techniques; see Chen et al. (2000) or Han & Carlin (2001) for a deeper coverage, and Meng & Schilling (2002) for the most efficient approach.

6·2. *Approximating Z by a formal reversible jump*

We first recover Gelfand & Dey's (1994) solution of reverse importance sampling by an integrated reversible jump, because a natural approach to compute a marginal likelihood is to

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

use a reversible jump Markov chain Monte Carlo algorithm (Green, 1995). We can indeed start from a single model $\mathcal{M}$ and still implement reversible jump in the following way. Consider a formal alternative model $\mathcal{M}'$, for instance a fixed distribution like the $\mathcal{N}(0, 1)$ distribution, with prior weight $1/2$ and build a proposal from $\mathcal{M}$ to $\mathcal{M}'$ that moves to $\mathcal{M}'$ with probability (Green, 1995) $\{(1/2)g(\theta)\}/\{(1/2)\pi(\theta)L(\theta)\} \wedge 1$, where $a \wedge b$ denotes $\min(a, b)$, and from $\mathcal{M}'$ to $\mathcal{M}$ with probability $\{(1/2)\pi(\theta)L(\theta)\}/\{(1/2)g(\theta)\} \wedge 1$, $g(\theta)$ being an arbitrary proposal on θ. Were we to actually run this reversible jump Markov chain Monte Carlo algorithm, the frequency of visits to $\mathcal{M}$ would then converge to Z.

However, the reversible sampler is not needed since, if we run a standard Markov chain Monte Carlo algorithm on θ and compute the probability of moving to M', the expectation of the ratio $g(\theta)/\pi(\theta)L(\theta)$ is equal to the inverse of Z:

$$E\{g(\theta)/\pi(\theta)L(\theta)\} = \int \frac{g(\theta)}{\pi(\theta)L(\theta)} \frac{\pi(\theta)L(\theta)}{Z} \, \mathrm{d}\theta = Z^{-1},$$

no matter what $g(\theta)$ is, in the spirit of both Gelfand & Dey (1994) and Bartolucci et al. (2006). This expression shows the connection of this approach with bridge sampling (Meng & Wong, 1996) since this also is a bridge sampling estimate where $\alpha(\theta) = 1/\pi(\theta)L(\theta)$.

Obviously, the choice of $g(\theta)$ affects the precision of the approximation to Z. When using a kernel approximation to $\pi(\theta \mid y)$ based on earlier Markov chain Monte Carlo simulations and considering the variance of the resulting estimator, the constraint is opposite to the one found in importance sampling, namely that $g(\theta)$ must have lighter, not fatter, tails than $\pi(\theta)L(\theta)$ for the approximation

$$\hat{Z}_1 = 1 \Big/ \left\{ \frac{1}{T} \sum_{t=1}^{T} g(\theta^{(t)})\pi(\theta^{(t)})L(\theta^{(t)}) \right\}$$

to have a finite variance. This means that light tails or finite support kernels, like an Epanechnikov kernel, are to be preferred to fatter tailed kernels, like the t kernel.

In the experimental comparison of § 7·2, we compare $\hat{Z}_1$ with a standard importance approximation

$$\hat{Z}_2 = \frac{1}{T} \sum_{t=1}^{T} \pi(\theta^{(t)})L(\theta^{(t)})/g(\theta^{(t)}), \quad \theta^{(t)} \sim g(\theta),$$

where g can also be a nonparametric approximation of $\pi(\theta \mid y)$, this time with heavier tails than $\pi(\theta)L(\theta)$. Frühwirth-Schnatter (2004) uses the same importance function g in both $\hat{Z}_1$ and $\hat{Z}_2$, and obtains results similar to ours, namely, that $\hat{Z}_2$ outperforms $\hat{Z}_1$.

6·3. *Approximating Z using a mixture representation*

Another approach in the approximation of Z is to design a specific mixture for simulation purposes, with density proportional to

$$m(\theta) \propto \omega_1 \pi(\theta)L(\theta) + g(\theta),$$

where $\omega_1 > 0$ and $g(\theta)$ is an arbitrary, fully specified density. Simulating from this mixture has the same complexity as simulating from the posterior, since the Markov chain Monte Carlo code used to simulate from $\pi(\theta \mid y)$ can be easily extended by introducing an auxiliary variable δ that indicates whether or not the current simulation is from $\pi(\theta \mid y)$ or from $g(\theta)$. The tth iteration of this extension is as follows, where $\mathcal{K}(\theta, \theta')$ denotes an arbitrary Markov kernel with stationary distribution the posterior $\pi(\theta \mid y) \propto \pi(\theta)L(\theta)$ as follows.

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

Step 1. Take $\delta^{(t)} = 1$, and $\delta^{(t)} = 2$ otherwise, with probability

$$\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) / \{\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) + g(\theta^{(t-1)})\}.$$

Step 2. If $\delta^{(t)} = 1$, generate $\theta^{(t)} \sim \mathcal{K}(\theta^{(t-1)}, \theta^{(t)})$, else generate $\theta^{(t)} \sim g(\theta)$ independently.

This algorithm is a Gibbs sampler: Step 1 simulates $\delta^{(t)}$ conditional on $\theta^{(t-1)}$, while Step 2 simulates $\theta^{(t)}$ conditional on $\delta^{(t)}$. While the average of the $\delta^{(t)}$s converges to $\omega_1 Z/\{\omega_1 Z + 1\}$, a natural Rao–Blackwellization is to take the average of the expectations of the $\delta^{(t)}$s,

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) / \{\omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + g(\theta^{(t)})\},$$

since its variance should be smaller. A third estimate is then deduced from this approximation by solving $\omega_1 \hat{Z}_3/\{\omega_1 \hat{Z}_3 + 1\} = \hat{\xi}$.

The use of mixtures in importance sampling in order to improve the stability of the estimators dates back at least to Hesterberg (1995) but, as it occurs, this particular mixture estimator happens to be almost identical to the bridge sampling estimator of Meng & Wong (1996). In fact,

$$\hat{Z}_3 = \frac{1}{\omega_1} \sum_{t=1}^{T} \frac{\omega_1 \pi(\theta^{(t)}) L(\theta^{(t)})}{\omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + g(\theta^{(t)})} \Big/ \sum_{t=1}^{T} \frac{g(\theta^{(t)})}{\omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + g(\theta^{(t)})}$$

is the Monte Carlo approximation to the ratio

$$E_m\{\alpha(\theta)\pi(\theta)L(y \mid \theta)\}/E_m\{\alpha(\theta)g(\theta)\}$$

when using the optimal function $\alpha(\theta) = 1/\{\omega_1 \pi(\theta) L(\theta) + g(\theta)\}$. The only difference with Meng & Wong (1996) is that, since the $\theta^{(t)}$s are simulated from the mixture, they can be recycled for both sums.

7. NUMERICAL EXPERIMENTS

7·1. *A decentred Gaussian example*

We modify the Gaussian toy example of § 4·2: $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})$, where the $\theta^{(k)}$s are independent and identically distributed from $\mathcal{N}(0, 1)$, and $y_k \mid \theta^{(k)} \sim \mathcal{N}(\theta^{(k)}, 1)$ independently, but setting all the y_ks to 3. To simulate from the prior truncated to $L(\theta) > L(\theta_0)$, we perform M Gibbs iterations with respect to this truncated distribution, with $M = 1, 3$ or 5: the full conditional distribution of $\theta^{(k)}$, conditional on $\theta^{(j)}$, $j \neq k$, is a $\mathcal{N}(0, 1)$ distribution that is truncated to the interval $[y^{(k)} - \delta, y^{(k)} + \delta]$ with

$$\delta^2 = \sum_j \left(y_j - \theta_0^{(j)}\right)^2 - \sum_{j \neq k} (y_j - \theta^{(j)})^2.$$

The nested sampling algorithm is run 20 times for $d = 10, 20, \ldots, 100$, and several combinations of (N, M): (100, 1), (100, 3), (100, 5) and (500, 1). The algorithm is stopped when a new contribution $(x_{i-1} - x_i)\varphi_i$ to (1) becomes smaller than 10^{-8} times the current estimate. Focusing first on $N = 100$, Fig. 1 exposes the impact of the mixing properties of the Markov chain Monte Carlo step: for $M = 1$ (a), the bias sharply increases with respect to the dimension, while, for $M = 3$ (b), it remains small for most dimensions. Results for $M = 3$ and $M = 5$ are quite similar, except perhaps for $d = 100$. Using $M = 3$ Gibbs steps seems to be sufficient to produce a good approximation of an ideal nested sampling algorithm, where points would be independently simulated. Interestingly, if N increases to 500, while keeping $M = 1$, then larger

This content downloaded from
81.111.181.125 on Thu, 1fff on Thu, 01 Jan 1976 12:34:56 UTC
All use subject to https://about.jstor.org/terms

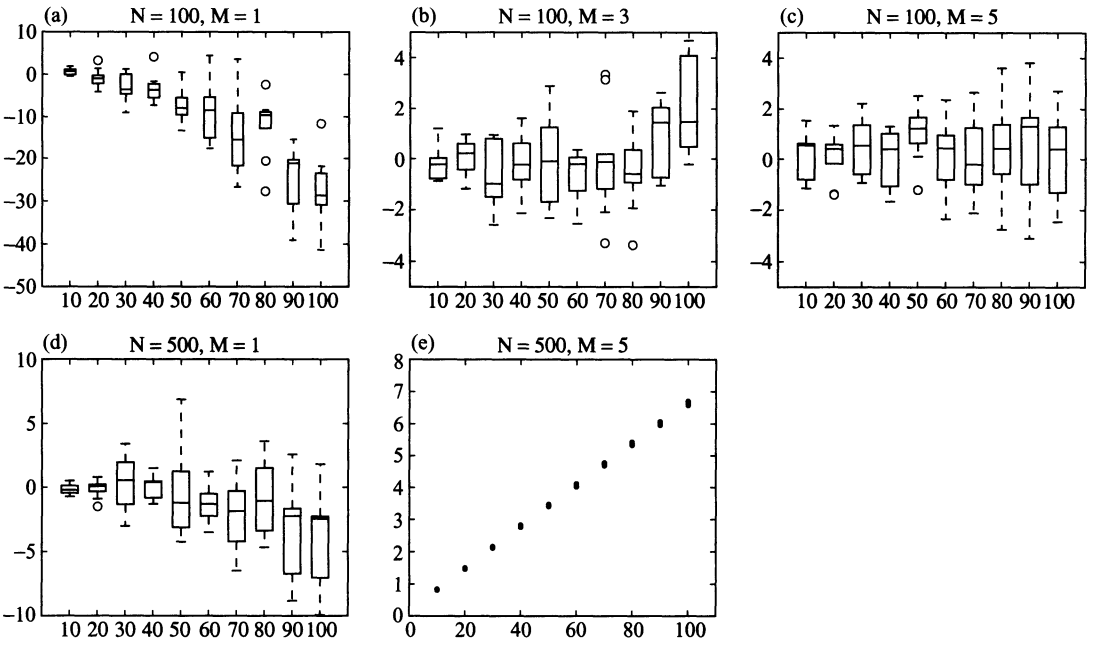

Fig. 1. Decentred Gaussian example: Box-plots of the log-relative error $\log \hat{Z} - \log Z$ versus dimension d for four values of (N, M), and (lower right) total number of iterations ($\times 10^4$) versus dimension for $(N, M) = (100, 5)$.

errors occur for the same computational effort, see Fig. 1 (d). Thus, a good strategy in this case is first to increase M until the distribution of the error stabilizes, then to increase N to reduce the Monte Carlo error. As expected, the number of iterations increases linearly with the dimension.

While artificial, this example shows that nested sampling may perform quite well even in large dimensional problems, provided M is large enough.

7·2. *A mixture example*

As in Frühwirth-Schnatter (2004), we consider the example of the posterior distribution on (μ, σ) associated with the normal mixture

$$y_1, \ldots, y_n \sim p\mathcal{N}(0, 1) + (1 - p)\mathcal{N}(\mu, \sigma), \tag{4}$$

when p is known, for two reasons. First, when σ goes to 0 and μ is equal to any of the x_is ($i = 1, \ldots, n$), the likelihood diverges; see Fig. 2. This is a priori challenging for exploratory schemes such as nested sampling. Second, efficient Markov chain Monte Carlo strategies have been developed for mixture models (Diebolt & Robert, 1994; Richardson & Green, 1997; Celeux et al., 2000), but Bayes factors are difficult to approximate in this setting.

We simulate n observations from a $\mathcal{N}\{2, (3/2)^2\}$ distribution, and then compute the estimates of Z introduced above for the model (4). The prior distribution is uniform on $(-2, 6) \times (0{\cdot}001, 16)$ for $(\mu, \log \sigma^2)$. The prior is arbitrary, but it allows for an easy implementation of nested sampling since the constrained simulation can be implemented via a random walk move.

The two-dimensional nature of the parameter space allows for a numerical integration of $L(\theta)$, based on a Riemann approximation and a grid of 800 × 500 points in the $(-2, 6) \times (0{\cdot}001, 16)$ square. This approach leads to a stable evaluation of Z that can be taken as the reference against which we can test the various methods, since additional evaluations based on a crude Monte Carlo integration using 10^6 terms and on Chib's method (1995) produced essentially the same

This content downloaded from
81.ffff:ffff:ffff:ffff:ffff on Thu, 01 Jan 1976 12:34:56 UTC
All use subject to https://about.jstor.org/terms

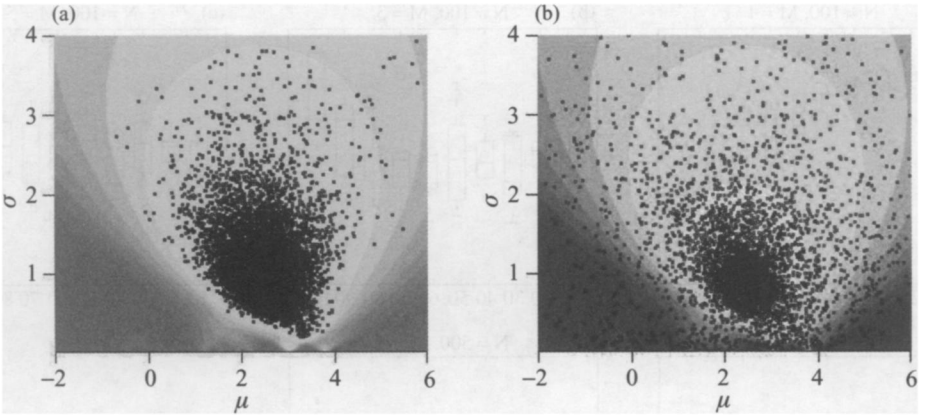

Fig. 2. Mixture example: (a) Markov chain Monte Carlo sample plotted on the loglikelihood surface in the (μ, σ) space for $n = 10$ observations from (4); (b) nested sampling sequence based on $N = 10^3$ starting points.

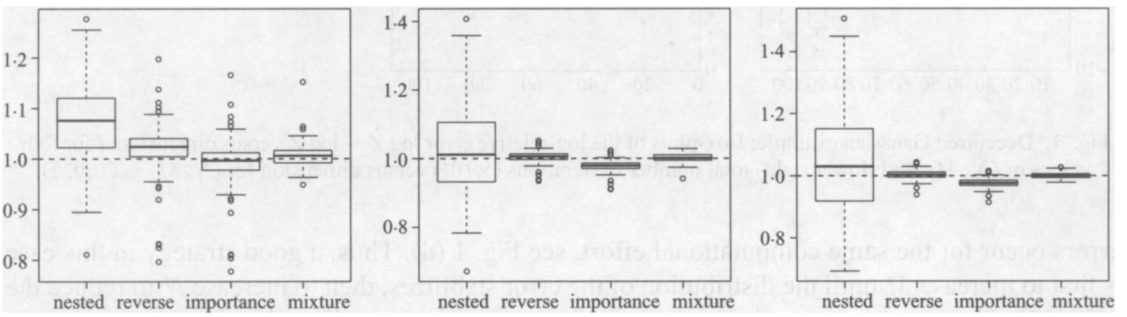

Fig. 3. Mixture model: comparison of the variations of nested sampling, reverse importance sampling, importance sampling and mixture sampling, relative to a numerical approximation of Z (dotted line), based on 150 samples of size $n = 10, 50, 100$.

numerical values. The Markov chain Monte Carlo algorithm implemented here is the standard completion of Diebolt & Robert (1994), but it does not suffer from the usual label switching deficiency (Jasra et al., 2005) because (4) is identifiable. As shown by the Markov chain Monte Carlo sample of size $N = 10^4$ displayed in Fig. 2(a), the exploration of the modal region by the Markov chain Monte Carlo chain is satisfactory. This Markov chain Monte Carlo sample is used to compute the nonparametric approximations g that appear in the three alternatives of § 6. For the reverse importance sampling estimate Z_1, g is a product of two Gaussian kernels with a bandwidth equal to half the default bandwidth of the R function density(), while, for both Z_2 and Z_3, g is a product of two t kernels with a bandwidth equal to twice the default Gaussian bandwidth.

We ran the nested sampling algorithm, with $N = 10^3$, reproducing the implementation of Skilling (2006), with 10 steps of a random walk in $(\mu, \log\sigma)$ constrained by the likelihood boundary. The total number of points produced by nested sampling at stopping time is on average close to 10^4, which justifies using the same number of points for the Markov chain Monte Carlo algorithm. As shown in Fig. 2(b), the nested sampling sequence visits minor modes of the likelihood surface but ends up in the same central mode as the Markov chain Monte Carlo sequence. All points visited by nested sampling are represented without reweighting, which explains for a larger density of points outside the central modal region.

The analysis of this Monte Carlo experiment in Fig. 3 shows that nested sampling gives approximately the same numerical value as the three other approaches, exhibiting a slight upward

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

bias, but that its variability is higher. The most reliable approach, besides the numerical and raw Monte Carlo evaluations that cannot be used in general settings, is the importance sampling solution, followed very closely by the mixture approach of § 6·3. The reverse importance sampling naturally shows a slight upward bias for the smaller values of n and a variability that is very close to both other alternatives, especially for larger values of n.

7·3. *A probit example for nested importance sampling*

To implement the nested importance sampling algorithm based on nested ellipsoids, we consider the arsenic dataset and a probit model studied in Chapter 5 of Gelman & Hill (2006). The observations are independent Bernoulli variables y_i such that $\text{pr}(y_i = 1 \mid x_i) = \Phi(x_i^{\mathrm{T}}\theta)$, where x_i is a vector of d covariates, θ is a vector parameter of size d and Φ denotes the standard normal distribution function. In this particular example, $d = 7$; more details on the data and the covariates are available on the book's web page http://www.stat.columbia.edu/∼ gelman/arm/examples/arsenic.

The probit model we use is model 9a in the R program (R Development Core Team, 2010) and is available at the above address: the dependent variable indicates whether or not the surveyed individual changed the well she drinks from over the past three years, and the seven covariates are an intercept, distance to the nearest safe well, in 100 m unit, education level, log of arsenic level and cross-effects for these three variables. We assign $\mathcal{N}_d(0, 10^2 I_d)$ as our prior on θ, and denote by θ_m the posterior mode, and Σ_m the inverse of minus twice the Hessian at the mode; both quantities are obtained numerically beforehand.

We ran the nested ellipsoid algorithm 50 times, for $N = 2$, 8, 32, 128, and for two sets of hyperparameters corresponding to both scenarios described in § 5. In the first scenario, $(\hat{\theta}, \hat{\Sigma}) = (\theta_m, 2\Sigma_m)$. Figures 4(c) and (d) compare log-errors produced by our method (c), with those of importance sampling based on the optimal Gaussian proposal, with mean θ_m, variance Σ_m and the same number of likelihood evaluations, as reported on the x-axis of (d). In the second scenario, $(\hat{\theta}, \hat{\Sigma}) = (\theta_m, 100\, I_d)$. Figures 4(a) and (b) compare log-errors produced by our method (a) with those of importance sampling, based again on the optimal proposal, and the same number of likelihood evaluations. The variance of importance sampling estimates based on a Gaussian proposal with hyperparameters $\hat{\theta}$ and $\hat{\Sigma} = 100 I_d$ is higher by several order of magnitudes, and is not reported in the plots.

As expected, the first strategy outperforms standard importance sampling, when both methods are supplied with the same information, and the second strategy still does reasonably well compared to importance sampling based on the optimal Gaussian proposal, although provided with the mode only. Results are sufficiently precise that evidence can be computed for all 2^7 possible models: the most likely model, with posterior probability 0·81, includes the intercept, the three variables mentioned above, distance, arsenic, education and one cross-effect between distance and education level, and the second most likely model, with posterior probability 0·18, simply removes the cross-effect.

8. Discussion

Nested sampling is a valid addition to the Monte Carlo toolbox, with convergence rate $O(N^{-1/2})$, and computational cost $O(d^3)$, where d is the dimension of the problem. It enjoys good performance in some applications, for example when the posterior is approximately Gaussian, but it may require more iterations to achieve the same precision in certain situations. Therefore, further work on nested sampling is needed. For one thing, the convergence properties of Markov chain Monte Carlo-based nested sampling are unknown and technically challenging. Methodologically, efforts are required to design efficient Markov chain Monte Carlo moves with

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

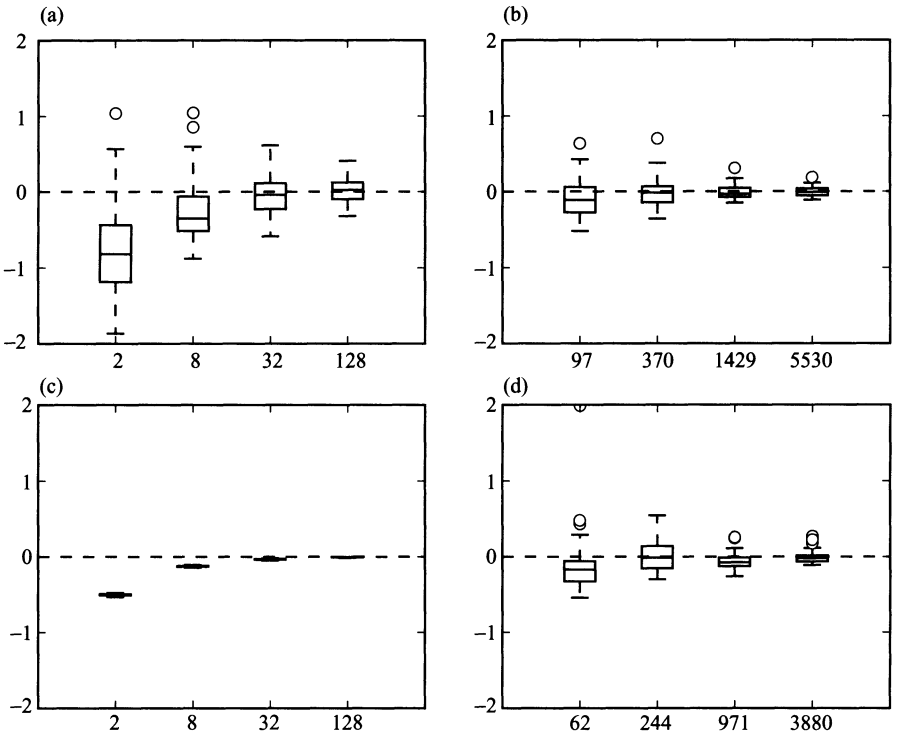

Fig. 4. Probit example: Box-plots of log-errors of nested importance sampling estimates (a and c), for $N = 2, 8, 32, 128$, compared with the log-error of importance sampling estimates based on the optimal Gaussian proposal, and the same number of likelihood evaluations (b and d). Those are reported on the x-axis of (b) and (d). Panels (c) and (d) correspond to the first strategy, based on both mode and Hessian, while (a) and (b) correspond to the second strategy, only based on mode.

respect to the constrained prior. In that and other respects, nested importance sampling may constitute a useful extension. Ultimately, our comparison between nested sampling and alternatives should be extended to more diverse examples in order to get a clearer idea of when nested sampling should be the method of choice and when it should not. For instance, Murray et al. (2006) report that nested sampling strongly outperforms annealed importance sampling (Neal, 2001) for Potts models. All the programs implemented for this paper are available from the authors.

Another discussion of the convergence of nested sampling may be found in Skilling (2009).

ACKNOWLEDGEMENT

The authors are grateful to R. Denny, A. Doucet, T. Loredo, O. Papaspiliopoulos, G. Roberts, J. Skilling, the editor, the associate editor and the referees for helpful comments. The authors are members of the Center for Research in Economics and Statistics. This work was partly supported by the Agence Nationale de la Recherche.

APPENDIX

Proof of Lemma 1. It is sufficient to prove the result for functions $\tilde{f}$ that are real-valued, positive and increasing. First, the extension to vector-valued functions is trivial. Second, the class of functions

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

that satisfy property (3) are clearly stable through addition. Since $\tilde{f}$ is absolutely continuous, there exist functions f^+ and f^-, such that f^+ is increasing, f^- is decreasing and $\tilde{f} = f^+ + f^-$, so we can restrict our attention to increasing functions. Third, absolute continuity implies bounded variation, so it is always possible to add an arbitrary constant to $\tilde{f}$ to transform it into a positive function.

To conclude the proof, take $\psi : l \to l\tilde{f}(l)$ to be a positive, increasing function with inverse ψ^{-1}; then

$$E^{\pi}[\psi\{L(\theta)\}] = \int_0^{+\infty} \mathrm{pr}[\psi\{L(\theta)\} > l]\,\mathrm{d}l = \int_0^{+\infty} \varphi^{-1}\{\psi^{-1}(l)\}\,\mathrm{d}l = \int_0^1 \psi\{\varphi(x)\}\,\mathrm{d}x. \qquad \square$$

Proof of Theorem 1. Let $t_i = x_{i+1}^{\star}/x_i^{\star}$, for $i = 0, 1, \ldots$. As mentioned by Skilling (2006), the t_is are independent Be(N, 1) variates. Thus, $u_i = t_i^N$ defines a sequence of independent uniform [0, 1] variates. A Taylor expansion of η_N gives

$$\begin{aligned}\eta_N &= \sum_{i=1}^{\lceil cN\rceil}(x_{i-1} - x_i)\{\varphi(x_i^{\star}) - \varphi(x_i)\}\\ &= \sum_{i=1}^{\lceil cN\rceil}(x_{i-1} - x_i)\{\psi'(-\log x_i)(\log x_i - \log x_i^{\star}) + O(\log x_i - \log x_i^{\star})^2\},\end{aligned}$$

where $c = -\log\varepsilon$ and $\psi(y) = \varphi(e^{-y})$. Furthermore,

$$S_i = N(\log x_i - \log x_i^{\star}) = \sum_{k=0}^{i-1}(-1 - \log u_k)$$

is a sum of independent, standard variables, as $E(\log u_i) = -1$ and $\mathrm{var}(\log u_i) = 1$. Thus, $(\log x_i - \log x_i^{\star}) = O_P(N^{-1/2})$, where the implicit constant in $O_P(N^{-1/2})$ does not depend on i, and

$$\begin{aligned}N^{1/2}\eta_N &= N^{-1/2}\sum_{i=1}^{\lceil cN\rceil}\left(e^{-(i-1)/N} - e^{-i/N}\right)S_i\left\{\psi'\left(\frac{i}{N}\right) + O_P\left(N^{-1/2}\right)\right\}\\ &= c^{1/2}\sum_{i=1}^{\lceil cN\rceil}\int_{(i-1)/N}^{i/N} e^{-t}\psi'(t)B_N\left(\frac{t}{c}\right)\mathrm{d}t\left\{1 + O_P\left(N^{-1/2}\right)\right\},\end{aligned}$$

since $\psi'(t) = \psi'(i/N) + O(N^{-1})$ for $t \in [(i-1)/N, i/N]$, where, again, the implicit constant in $O(N^{-1})$ can be the same for all i, as ψ'' is bounded, and provided $B_N(t)$ is defined as $B_N(t) = (cN)^{-1/2}S_{\lceil cNt\rceil}$ for $t \in [0, 1]$. According to Donsker's theorem (Kallenberg, 2002, p. 275), B_N converges to a Brownian motion B on [0, 1], in the sense that $f(B_N)$ converges in distribution to $f(B)$ for any measurable and almost surely continuous function f. Thus

$$N^{1/2}\eta_N = c^{1/2}\int_0^{\lceil cN\rceil/N} e^{-t}\psi'(t)B_N\left(\frac{t}{c}\right)\mathrm{d}t + O_P\left(N^{-1/2}\right)$$

converges in distribution to

$$c^{1/2}\int_0^c e^{-t}\psi'(t)B\left(\frac{t}{c}\right)\mathrm{d}t,$$

which has the same distribution as the following zero-mean Gaussian variate:

$$\int_0^c e^{-t}\psi'(t)B(t)\,\mathrm{d}t = \int_{\varepsilon}^1 s\varphi'(s)B(-\log s)\,\mathrm{d}s. \qquad \square$$

Proof of Lemma 2. For the sake of clarity, we make dependencies on d explicit in this section, including φ_d for φ, ε_d for ε and so on. We will use repeatedly the facts that φ is nonincreasing and that φ' is

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

nonnegative. One has

$$-\int_{s,t\in[\varepsilon_d,1]} s\varphi_d'(s)t\varphi_d'(t)\log(s\vee t)\,\mathrm{d}t \leqslant -\log\varepsilon_d\left\{\int_{\varepsilon_d}^1 s\varphi_d'(s)\,\mathrm{d}s\right\}^2 \leqslant d\log\left(2^{1/2}/\tau\right)$$

for $d \geqslant 1$, since $-\int_{\varepsilon_d}^1 s\varphi_d'(s)\,\mathrm{d}s \leqslant -\int_0^1 s\varphi_d'(s)\,\mathrm{d}s = 1$. This gives the first result.

Let $s_d = \varphi_d^{-1}(\alpha^d)$, for $0 < \alpha < 1$; s_d is the probability that

$$(4\pi/d)\sum_{i=1}^d \theta_i^2 - 1 \leqslant -2\log(\alpha) + \log(2) - 1$$

assuming that the θ_is are independent $\mathcal{N}(0, 1/4\pi)$ variates. The left-hand side is an empirical average of independent and identically distributed zero-mean variables. We take α so that the right-hand side is negative, which implies $\alpha > 2^{1/2}\exp(-1/2)$. Using large deviations (Kallenberg, 2002, Ch. 27), one has $-\log(s_d)/d \to \gamma > 0$ as $d \to +\infty$, and

$$\begin{aligned}
\frac{1}{d}V_d &= -\frac{1}{d}\int_{s,t\in[\varepsilon_d,1]} s\varphi_d'(s)t\varphi_d'(t)\log(s\vee t)\,\mathrm{d}s\,\mathrm{d}t \\
&\geqslant \left(\frac{-\log s_d}{d}\right)\left\{\int_{\varepsilon_d}^{s_d} s\varphi_d'(s)\,\mathrm{d}s\right\}^2 \\
&\geqslant \left(\frac{-\log s_d}{d}\right)\left\{\int_{\varepsilon_d}^{s_d} \varphi_d(s)\,\mathrm{d}s + \varepsilon_d\varphi_d(\varepsilon_d) - s_d\varphi_d(s_d)\right\}^2 \\
&\geqslant \left(\frac{-\log s_d}{d}\right)\left\{1 - \int_0^{\varepsilon_d}\varphi_d(s)\,\mathrm{d}s - \int_{s_d}^1 \varphi_d(s)\,\mathrm{d}s + \varepsilon_d\varphi_d(\varepsilon_d) - s_d\varphi_d(s_d)\right\}^2.
\end{aligned}$$

As $d \to +\infty$, $-\log(s_d)/d \to \gamma$, $s_d \to 0$, $\varphi_d(s_d) = \alpha^d \to 0$, $\int_{s_d}^1 \varphi_d(s)\,\mathrm{d}s \leqslant \varphi_d(s_d)(1 - s_d) \to 0$, and

$$0 \leqslant \int_0^{\varepsilon_d}\varphi_d(s)\,\mathrm{d}s - \varepsilon_d\varphi_d(\varepsilon_d) \leqslant \varepsilon_d\{\varphi_d(0) - \varphi_d(\varepsilon_d)\} \leqslant \tau < 1,$$

by the definition of ε_d, and the squared factor is in the limit greater than or equal to $(1 - \tau)^2$. □

REFERENCES

BARTOLUCCI, F., SCACCIA, L. & MIRA, A. (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika* **93**, 41–52.

CELEUX, G., HURN, M. & ROBERT, C. (2000). Computational and inferential difficulties with mixtures posterior distribution. *J. Am. Statist. Assoc.* **95**, 957–79.

CHEN, M. & SHAO, Q. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25**, 1563–94.

CHEN, M., SHAO, Q. & IBRAHIM, J. (2000). *Monte Carlo Methods in Bayesian Computation*. New York; Springer.

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Am. Statist. Assoc.* **90**, 1313–21.

DIEBOLT, J. & ROBERT, C. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. R. Statist. Soc.* B **56**, 363–75.

EVANS, M. (2007). Discussion of Nested sampling for Bayesian computations by John Skilling. In *Bayesian Statistics* 8, Ed. J. Bernardo, M. Bayarri, J. Berger, A. David, D. Heckerman, A. Smith and M. West, pp. 491–524. Oxford: Oxford University Press.

FRÜHWIRTH-SCHNATTER, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometr. J.* **7**, 143–67.

GELFAND, A. & DEY, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc.* B **56**, 501–14.

GELMAN, A. & HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

GREEN, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–32.

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms

HAN, C. & CARLIN, B. (2001). MCMC methods for computing Bayes factors: a comparative review. *J. Am. Statist. Assoc.* **96**, 1122–32.

HESTERBERG, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 185–94.

JASRA, A., HOLMES, C. & STEPHENS, D. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20**, 50–67.

JEFFREYS, H. (1939). *Theory of Probability*, 1st ed. Oxford: The Clarendon Press.

KALLENBERG, O. (2002). *Foundations of Modern Probability*. New York: Springer.

LOUIS, T. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B **44**, 226–33.

MACLACHLAN, G. & KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley.

MENG, X. & SCHILLING, S. (2002). Warp bridge sampling. *J. Comp. Graph. Statist.* **11**, 552–86.

MENG, X. & WONG, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6**, 831–60.

MUKHERJEE, P., PARKINSON, D. & LIDDLE, A. (2006). A nested sampling algorithm for cosmological model selection. *Astrophys. J.* **638**, L51–L54.

MURRAY, I., MACKAY, D. J., GHAHRAMANI, Z. & SKILLING, J. (2006). Nested sampling for Potts models. In *Advances in Neural Information Processing Systems* 18, Ed. Y. Weiss, B. Schölkopf and J. Platt. Cambridge, MA: MIT Press.

NEAL, R. (2001). Annealed importance sampling. *Statist. Comp.* **11**, 125–39.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. URL: http://www.R-project.org.

RICHARDSON, S. & GREEN, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B **59**, 731–92.

ROBERT, C. (2001). *The Bayesian Choice*, 2nd ed. New York: Springer.

ROBERTS, G. & ROSENTHAL, J. (1999). Convergence of slice sampler Markov chains. *J. R. Statist. Soc.* B **61**, 643–60.

RYDÉN, T. (1994). Parameter estimation for Markov modulated Poisson processes. *Stochastic Models* **10**, 795–829.

SHAW, J., BRIDGES, M. & HOBSON, M. (2007). Efficient Bayesian inference for multimodal problems in cosmology. *Mon. Not. R. Astron. Soc.* **378**, 1365–70.

SKILLING, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Anal.* **1**, 833–60.

SKILLING, J. (2009). Nested sampling's convergence. In *AIP Proc.* 1193, pp. 277–91. New York: AIP.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B **64**, 583–639.

VEGETTI, S. & KOOPMANS, L. V. E. (2009). Bayesian strong gravitational-lens modelling on adaptive grids: objective detection of mass substructure in galaxies. *Mon. Not. R. Astron. Soc.* **392**, 945–63.

[*Received October* 2008. *Revised December* 2009]

This content downloaded from
81.111.181.125 on Thu, 15 Oct 2020 11:48:37 UTC
All use subject to https://about.jstor.org/terms