Evaluation of marginal likelihoods via the density of states

Michael Habeck
Department of Protein Evolution
Max Planck Institute for Developmental Biology
Spemannstrasse 35,
72076 Tübingen, Germany

Department of Empirical Inference
Max Planck Institute for Intelligent Systems
Spemannstrasse 38
72076 Tübingen, Germany

Abstract

Bayesian model comparison involves the evaluation of the marginal likelihood, the expectation of the likelihood under the prior distribution. Typically, this high-dimensional integral over all model parameters is approximated using Markov chain Monte Carlo methods. Thermodynamic integration is a popular method to estimate the marginal likelihood by using samples from annealed posteriors. Here we show that there exists a robust and flexible alternative. The new method estimates the density of states, which counts the number of states associated with a particular value of the likelihood. If the density of states is known, computation of the marginal likelihood reduces to a one-dimensional integral. We outline a maximum likelihood procedure to estimate the density of states from annealed posterior samples. We apply our method to various likelihoods and show that it is superior to thermodynamic integration in that it is more flexible with regard to the annealing schedule and the family of bridging distributions. Finally, we discuss the relation of our method with Skilling's nested sampling.

1 Introduction

Computation of the marginal likelihood – also called the evidence (MacKay, 2003) – is a crucial step in Bayesian model comparison. If our model M is based on the prior $\pi(\theta) \equiv \Pr(\theta|M)$ and the likelihood $L(\theta) \equiv \Pr(D|\theta, M)$ where θ are the parameters of the model and D are the data, the marginal likelihood is given by the integral

$$Z = \int L(\theta)\, \pi(\theta)\, \mathrm{d}\theta. \tag{1}$$

Bayes factors, that is ratios of marginal likelihoods or rather the difference of their logarithms, allow us to rank two competing models relative to each other and to evaluate how much each of them is supported by the data (Jeffreys, 1939; Kass and Raftery, 1995).

Calculation of the marginal likelihood is very challenging and is analytically intractable for most models of interest (Gelman and Meng, 1998). Standard numerical quadrature schemes fail to work due to the curse of dimensionality. A remedy is offered by Markov chain Monte Carlo (MCMC) methods. But if the posterior probability, $p(\theta) = L(\theta)\, \pi(\theta)/Z$, is multimodal and the parameters are highly correlated, sampling from the posterior itself is a nontrivial task. For complex posterior distributions, MCMC methods based on a family of distributions that interpolate between the prior and the posterior have proven to be successful (Gelman and Meng, 1998). Often the family of distributions is constructed by introducing a fictitious temperature and samples are drawn from annealed or power posteriors (Friel and Pettitt, 2008). An instance of thermal sampling is annealed importance sampling (AIS) (Neal, 2001), which can be viewed as an MCMC version of simulated annealing. Parallel tempering or replica exchange Monte Carlo (Swendsen and Wang, 1986; Geyer, 1991) is a population version of AIS, in which samples at all temperatures are generated in parallel and exchanged.

The annealed posterior is defined as

$$p(\theta|\beta) = L(\theta)^\beta\, \pi(\theta)/c(\beta) \tag{2}$$

where β is the inverse temperature and the normalization constant is

$$c(\beta) = \int L(\theta)^\beta\, \pi(\theta)\, \mathrm{d}\theta. \tag{3}$$

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

The marginal likelihood is only well-defined if the prior is normalizable ($c(0) < \infty$) and amounts to $Z = c(1)/c(0)$. The standard approach to calculate the evidence from annealed posterior samples is to use thermodynamic integration (TI) (Kirkwood, 1935; Gelman and Meng, 1998). TI is based on the observation that

$$\begin{aligned}\log\{c(1)/c(0)\} &= \int_0^1 \frac{\mathrm{d}\log c(\beta)}{\mathrm{d}\beta}\,\mathrm{d}\beta \\ &= \int_0^1 \left[\int \log L(\theta)\,\frac{L(\theta)^\beta \pi(\theta)}{c(\beta)}\,\mathrm{d}\theta\right]\mathrm{d}\beta \\ &= \int_0^1 \langle \log L\rangle_\beta\,\mathrm{d}\beta \qquad (4)\end{aligned}$$

where $\langle\,\cdot\,\rangle_\beta$ indicates the expectation with respect to the annealed posterior at inverse temperature β. If we replace the β-integral with a sum over average log-likelihood values at discrete inverse temperatures $\beta_1, \ldots, \beta_m$, we obtain an approximate value of the log-evidence:

$$\log Z \approx \frac{1}{2}\sum_{i=1}^{m-1}(\beta_{i+1}-\beta_i)(\overline{\log L}_{i+1} + \overline{\log L}_i) \qquad (5)$$

where $\overline{\log L}_i$ is the arithmetic average of log-likelihood values calculated over all samples drawn from the ith bridging distribution.

An alternative to thermodynamic integration is to focus on the density of states and to consider it the primary object of interest from which the evidence follows as a secondary quantity. The density of states (DOS) is defined as

$$g(E) = \int \delta(E + \log L(\theta))\,\pi(\theta)\,\mathrm{d}\theta \qquad (6)$$

and measures the prior mass associated with a particular value of the energy $E(\theta) = -\log L(\theta)$. If the density of states is known, evaluation of the marginal likelihood reduces to the one-dimensional integral

$$\begin{aligned}Z &= \int L(\theta)\,\pi(\theta)\,\mathrm{d}\theta \\ &= \int\left[\int \delta(E+\log L(\theta))\,\pi(\theta)\,\mathrm{d}\theta\right] e^{-E}\,\mathrm{d}E \\ &= \int g(E)\,e^{-E}\,\mathrm{d}E. \qquad (7)\end{aligned}$$

In the next section, we outline how to estimate $g(E)$ from MCMC samples and how to compute the evidence using our estimate of the DOS. We then show for various examples that DOS based evidence estimation has advantages over thermodynamic integration. Finally, we discuss the relation of the DOS method with nested sampling (Skilling, 2004; Sivia and Skilling, 2006).

2 Non-parametric estimation of the density of states

We assume that samples were generated from a series of distributions

$$\theta_i \sim q_i(E(\theta))\,\pi(\theta)/c_i, \quad i = 1, \ldots, m \qquad (8)$$

where $c_i = \int q_i(E(\theta))\,\pi(\theta)\,\mathrm{d}\theta = \int g(E)\,q_i(E)\,\mathrm{d}E$ and q_i is a family of non-negative functions or *ensembles*. For $q_i(E) = e^{-\beta_i E}$ we obtain the annealed posterior (2) at inverse temperatures β_i. In bridge sampling, the initial and final ensemble satisfy $q_1(E) = 1$ and $q_m(E) = e^{-E}$ such that family (8) morphs the prior into the posterior. However, we are free to choose any other family of distributions that allows us to focus on the high posterior probability regions. In case of $q_i(E) = \Theta(E_i - E)$ where $\Theta(\cdot)$ is the Heaviside step function and $E_{i+1} < E_i$, we sample from the prior under the likelihood constraints $L(\theta) \geq e^{-E_i}$; this choice is relevant to nested sampling.

2.1 Maximum likelihood estimation of the density of states

We use a maximum likelihood approach to estimate the density of states $g(E)$. Let $E_i = -\log L(\theta_i)$ be the negative log-likelihoods of our samples. The likelihood of generating E_i from the ith ensemble is

$$E_i \sim g(E)\,q_i(E)/c_i.$$

It is more convenient to estimate the logarithm of the DOS $u(E) = \log g(E)$ rather than the DOS itself. Maximum likelihood then optimizes over the space of all square-integrable functions, which is equipped with the inner product $\langle u, v\rangle = \int u(E)\,v(E)\,\mathrm{d}E$.

Because c_i depends on u but q_j doesn't, we need to minimize the functional

$$\begin{aligned}\mathcal{L}[u] &= -\sum_i u(E_i) + \sum_i \log\left\{\int q_i(E)\,e^{u(E)}\,\mathrm{d}E\right\} \\ &= -\langle h, u\rangle + \sum_i \log c_i[u] \qquad (9)\end{aligned}$$

where we introduced the empirical energy histogram $h(E) = \sum_i \delta(E - E_i)$ with an infinitely fine binning. $\mathcal{L}[u]$ is convex in u, which follows from the log-convexity of the normalization constants. For two functions u, v and scalar $\alpha \in [0,1]$, we have

$$\begin{aligned}&\log c_i[\alpha\,u + (1-\alpha)\,v] \\ &\qquad = \log\left\{\int [q_i\,e^u]^\alpha\,[q_i\,e^v]^{(1-\alpha)}\right\} \\ &\qquad \leq \alpha\,\log c_i[u] + (1-\alpha)\,\log c_i[v]\end{aligned}$$

by application of Hölder's inequality. Because $\mathcal{L}[u]$ is convex, we can determine u uniquely up to translation (addition of a constant to u does not change $\mathcal{L}[u]$) by solving

$$\frac{\delta\mathcal{L}[u]}{\delta u} = -h + \sum_i \frac{1}{c_i[u]}\, q_i\, e^u = 0.$$

This non-linear functional equation cannot be solved explicitly for u. We use a majorization-minimization strategy to optimize $\mathcal{L}[u]$. Instead of minimizing $\mathcal{L}[u]$ directly, we minimize an adaptive upper bounding functional $\tilde{\mathcal{L}}[u, \tilde{u}]$ that is derived at the current estimate $\tilde{u}$. We obtain the iteration:

$$u^{(t+1)} = \arg\min_u \tilde{\mathcal{L}}[u, u^{(t)}]. \tag{10}$$

If we choose $\tilde{\mathcal{L}}[u, \tilde{u}]$ tangent to $\mathcal{L}[u]$ at $\tilde{u}$, i.e. $\mathcal{L}[\tilde{u}] = \tilde{\mathcal{L}}[\tilde{u}, \tilde{u}]$, iteration (10) converges to the global minimum of $\mathcal{L}[u]$. To derive an upper bounding functional with the desired properties, we use $\log x \leq x - 1$ with equality at $x = 1$ to obtain:

$$\tilde{\mathcal{L}}[u, \tilde{u}] = -\langle h, u\rangle + \sum_i \log c_i[\tilde{u}] + \sum_i \left(\frac{c_i[u]}{c_i[\tilde{u}]} - 1\right). \tag{11}$$

From $\delta\tilde{\mathcal{L}}[u, \tilde{u}]/\delta u = 0$, we have:

$$-h + \sum_i \frac{1}{c_i[\tilde{u}]}\, q_i\, e^u = 0$$

and obtain an explicit representation of the optimal u at each iteration:

$$\exp u^{(t+1)} = \frac{h}{\sum_i q_i / c_i[u^{(t)}]}. \tag{12}$$

Update (12) implies a recursive relation for the normalization constants $c_j^{(t)} \equiv c_j[u^{(t)}]$:

$$c_j^{(t+1)} = \sum_i \frac{q_{ji}}{\sum_k q_{ki}/c_k^{(t)}} \tag{13}$$

where $q_{ji} = q_j(E_i)$. Using (13) it is straightforward to verify that indeed $\mathcal{L}[u^{(t+1)}] < \mathcal{L}[u^{(t)}]$. Therefore iteration (12) will lead us to the global minimum of the negative log-likelihood functional $\mathcal{L}[u]$.

Iteration (12) has been derived first by Ferrenberg and Swendsen (1989) for discrete thermodynamic systems such as Ising models. The energy histogram h is a sufficient statistic, which is why (12) is also called histogram reweighting. Histogram reweighting has been generalized to continuous systems resulting in the weighted histogram analysis method (WHAM) often used to analyze biomolecular simulations (Kumar *et al.*, 1992). One disadvantage of WHAM is that energies are discretized in order to work with an explicit energy histogram. This introduces some arbitrariness because the optimal binning has to be determined somehow. However, because we never need to work with an explicit representation of u but can rather update the normalization constants according to (13), there is no need to bin the energies and we achieve a truly non-parametric estimate of the density of states.

2.2 Estimation of the marginal likelihood

In iteration (12), we never really need to work with the explicit representation of u because of the recursive definition of the normalization constants. Therefore our algorithm to estimate DOS or its logarithm only involves iterating over (13). It is more convenient to work in log-space and update the free energies $f_j = -\log c_j$ rather than the normalization constants:

$$f_j^{(t+1)} = -\log \sum_i \frac{q_{ji}}{\sum_k q_{ki}\, e^{f_k^{(t)}}}, \quad f_j^{(0)} = 0. \tag{14}$$

This iteration is identical to the multistate Bennett acceptance ratio (Shirts and Chodera, 2008). The convergence of (14) is monitored by expressing $\mathcal{L}[u^{(t)}]$ as a function of the current estimate of the free energies:

$$\mathcal{L}(f^{(t)}) = -\sum_j f_j^{(t)} + \sum_i \log \sum_j q_{ji} e^{f_j^{(t)}} \tag{15}$$

where terms independent of $f^{(t)}$ are omitted. $\mathcal{L}(f)$ has the same functional form as $\mathcal{L}[u]$ (9), the only difference being that $\mathcal{L}(f)$ is defined on a finite dimensional vector space whereas $\mathcal{L}[u]$ operates on function space.

After convergence of iteration (14), our non-parametric estimate of the density of states is given by $\hat{g} = h/\sum_i e^{\hat{f}_i} q_i$ where $\hat{f}$ is the final result of (14). Finally, the DOS based evidence estimate is given by

$$\hat{Z} = \frac{\sum_i e^{-E_i}/\sum_j q_{ji}\, e^{\hat{f}_j}}{\sum_k 1/\sum_j q_{jk}\, e^{\hat{f}_j}} = \sum_i \hat{g}_i\, e^{-E_i} \tag{16}$$

where we introduced the estimated density of states at the sampled energies E_i

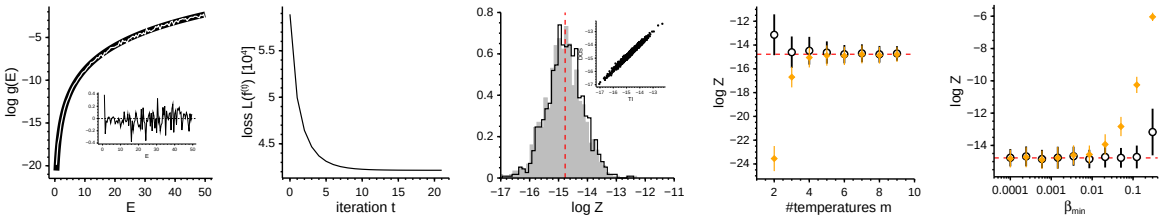
$$\hat{g}_i = \frac{(\sum_j q_{ji}\, e^{\hat{f}_j})^{-1}}{\sum_k (\sum_j q_{jk}\, e^{\hat{f}_j})^{-1}} \tag{17}$$

for convenience.

3 Applications and comparison to thermodynamic integration

3.1 Gaussian likelihood

As a first test, we consider the unnormalized Gaussian in d dimensions as likelihood. We use a flat prior and



(a) (b) (c) (d) (e)

Figure 1: Marginal likelihood estimation for the Gaussian likelihood over the 10-dimensional unit ball. (a): Estimated DOS (white line) and true DOS (black band). The inset shows the discrepancy of the estimated and true DOS. (b): Evolution of the loss (15) during DOS estimation. (c): Histograms of estimated evidences over 1000 repetitions ($m = 10, n = 10$) using TI (grey filled histogram) and DOS (black curve); the true evidence is marked with a dashed red line. The inset displays a scatter plot of TI based against DOS based evidence estimates and shows that both are highly correlated. (d): Effect of the number of temperatures on the accuracy of the evidence estimates. Orange diamonds and error bars indicate the average and standard deviation of TI based evidence estimates for different numbers of temperatures (100 repetitions). Black circles and error bars indicate the results for DOS based evidence estimation. The true evidence is shown as dashed red line. (e): Effect of the minimum inverse temperature $\beta_{\min}$ on the accuracy of the evidence estimates. Same symbols and coloring as in Fig. 1(d).

restrict the parameter space to the d-dimensional unit ball ($\|\theta\| \leq 1$):

$$L(\theta) = \exp\{-\lambda\,\theta^T\theta/2\}, \quad \pi(\theta) = \frac{d\,\Theta(1-\|\theta\|)}{S(d)} \tag{18}$$

where $S(d) = 2\,\pi^{d/2}/\Gamma(d/2)$; λ determines how densely θ is distributed about the origin. The evidence and the density of states are given by:

$$Z = (d/2)\,(2/\lambda)^{d/2}\,\gamma(d/2, \lambda/2) \tag{19}$$

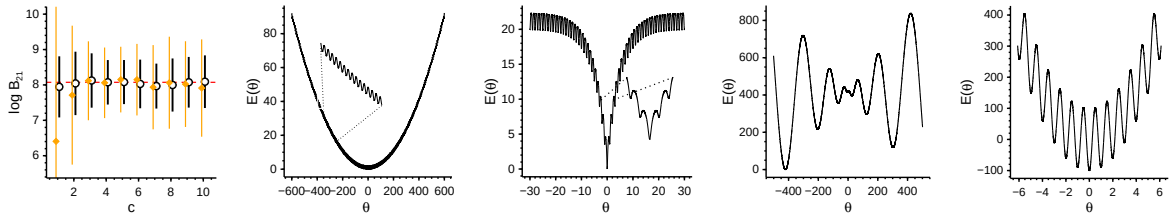
$$g(E) = (d/\lambda)\,(2E/\lambda)^{d/2-1}\,\Theta(\lambda/2 - E) \tag{20}$$

where $\gamma(s,x) = \int_0^x t^{s-1}\,e^{-t}\,\mathrm{d}t$ is the incomplete gamma function. We generate n samples at m equispaced inverse temperatures starting at $\beta_{\min} \geq 0$. Annealed posterior samples are generated in spherical coordinates $\theta = rx$ where x is a d-dimensional normalized direction ($\|x\| = 1$) and r the scalar radius ($r \leq 1$). x is drawn from a d-dimensional Gaussian and normalized to length 1. Samples of the radius are obtained by letting $r = \sqrt{2s}$ where s is generated from a Gamma distribution with shape $d/2$ and scale $(\beta\lambda)^{-1}$, restricted to the interval $[0, 1/2]$.

Figure 1 shows a comparison of estimated marginal likelihoods using thermodynamic integration (5) and the density of states (16). We choose $\lambda = 100$ and $d = 10$ and study the effect of varying the number of temperatures m as well as the choice of the minimum temperature $\beta_{\min}$. Figures 1(a) and 1(b) show that the estimated DOS is very close to the true DOS (19) and that the estimation quickly converges in a monotonic fashion. For sufficiently many temperatures and small $\beta_{\min} = 10^{-5}$, TI and DOS basically give the same result (see Figure 1(c)). However as we decrease the number of temperatures, the TI estimate drifts systematically away from the true value because the evidence integral (4) is no longer approximated accurately. The DOS based evidence estimate, on the other hand, is stable against decreasing the number of temperatures and even works for the extreme case $m = 2$. A similar effect is observed when increasing $\beta_{\min}$, which is relevant to parallel tempering where one aims to choose $\beta_{\min} > 0$ in order to reduce the number of replicas, as long as $p(\theta|\beta_{\min})$ can be sampled ergodically. Figure 1(e) shows that TI systematically over-estimates evidences because for too large $\beta_{\min}$ a significant contribution to the TI integral (5) is missing. Again, DOS based evidence estimation is robust against increasing $\beta_{\min}$.

3.2 Linear regression

Next, we fit a straight line to pairs of observations. In the radiata pine data analyzed by Friel and Pettitt (2008), two inputs x_i and z_i are tested in terms of their ability to explain output y_i. The priors of the slope, the intercept and the variance are conjugate such that we can straightforwardly apply Gibbs sampling to generate samples from the annealed posterior. At each temperature, 10 samples are stored after a burnin phase of 100 iterations. Friel and Pettitt test different annealing schedules of the form $\beta_i = [(i-1)/(m-1)]^c$ to calculate the Bayes factor B_{21} in favor of regression on z_i over regression on x_i. The evidences of both models can be calculated numerically after analytical marginalization over the slope and the intercept. Comparison of evidences and Bayes factors shows that the



(a) (b) (c) (d) (e)

Energy	true log-evidence	TI estimate	DOS estimate
Griewank	-3.13	-3.20 ± 0.05	-3.09 ± 0.06
Ackley	-5.53	-5.58 ± 0.06	-5.49 ± 0.05
Schwefel	-5.30	-6.71 ± 0.17	-5.24 ± 0.16
Rastrigin	94.45	93.41 ± 0.13	94.03 ± 0.13

(f)

Figure 2: (a): Comparison of two regression models for the radiata pine data. Orange diamonds and error bars indicate the average and standard deviation of TI based Bayes factor estimates for different exponents c of the annealing schedule (100 repetitions). Black circles and error bars indicate the results for DOS based evidence estimation. The correct Bayes factor is shown as dashed red line. (b-e): Rugged energy functions analyzed with thermodynamic integration and with the density of states (Griewank (b), Ackley (c), Schwefel (d) and Rastrigin (e) function). For the Griewank and Ackley function, a magnified region of the energy function is shown as inset. (f): Table summarizing the results of TI based and DOS based evidence estimation. True log-evidences are calculated using numerical integration.

DOS based evidence estimate is more accurate then TI especially if the annealing schedule is not optimal (see Figure 2(a)). Moreover, the DOS analysis shows smaller variances over 100 repetitions indicating that the Bayes factor estimate is systematically more accurate.

3.3 Evidence estimation for multimodal likelihoods using parallel tempering

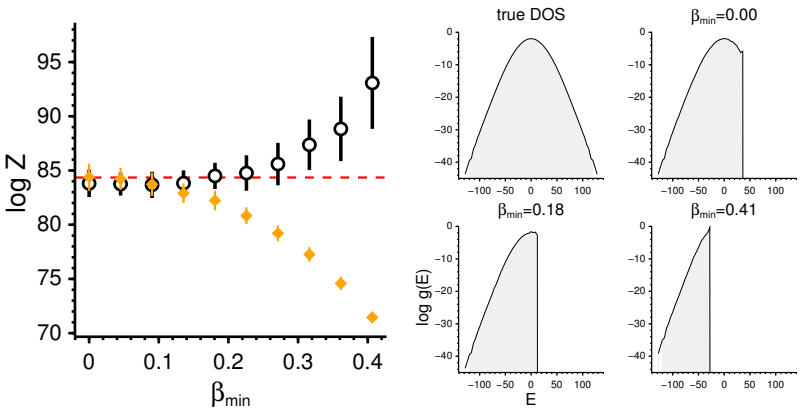
The previous examples involved unimodal posteriors. Let us now look at more challenging examples and calculate the evidence for some of the rugged likelihoods listed by Li *et al.* (2009) (shown in Figures 2(b)-2(e)). In all examples, the prior distribution is flat and bounded to a finite one-dimensional domain. Samples are generated using parallel tempering with 10 inverse temperatures between 0 and 1 following the same schedule as in the regression example (section 3.2) with exponent $c = 2$. 5000 exchange transitions are simulated, of which the first 2500 are considered as burnin. Samples from the annealed posteriors are generated with random walk Metropolis Monte Carlo. Only every 10th of the final 2500 samples is pooled into a single data set. This results in 10 different energy sets over which we calculate the mean and variance of the evidence estimates reported in Table 2(f). The results show that DOS based evidence estimation tends to be more accurate than thermodynamic integration. This is especially true for energy functions with deep minima such as the Schwefel function shown in Figure 2(d). For very rugged energy functions, the annealing schedule would require more tuning in order to achieve as accurate evidence estimates with thermodynamic integration as by using the density of states.

3.4 Two-dimensional Ising model

We also analyzed high-dimensional likelihoods to verify that the above findings also hold for more complex systems. The $L \times L$ Ising model is a discrete system with L^2 spins $\theta_i \in \{-1, +1\}$ that are either "up" or "down". The prior is flat and the likelihood function is given by:

$$L(\theta) = \exp\Big\{J \sum_{i \sim j} \theta_i \theta_j\Big\} \tag{21}$$

where $i \sim j$ indicates that spins i and j are nearest neighbors on a square lattice; J is the energy scale. In our tests, we set $J = 1$ and $L = 8$, the critical inverse temperature is $\beta_c \approx 0.4066$. The log-evidence (or free-energy), $\log Z = \log\{c(1)/c(0)\}$, of the system



(a) (b)

Energy function	TI estimate	DOS estimate
Purely repulsive energy	-963.1 ± 1.2	-960.7 ± 1.2
Repulsive/attractive energy	-846.9 ± 1.0	-845.8 ± 0.9

(c)

Figure 3: Comparison of evidence estimates for high-dimensional likelihoods. (a): 8×8 Ising model on a 2D square lattice. Log-evidences are calculated over 100 repetitions of thermal sampling at 11 temperatures. At each temperature, 100 samples were generated. Orange diamonds and error bars indicate the average and standard deviation of TI based evidence estimates for different $\beta_{\min}$. Black circles and error bars indicate the results for DOS based evidence estimation. The true evidence is shown as dashed red line. (b): Exact and estimated density of states for the 8×8 Ising model. (c): Model comparison for protein simulations. Two different physical energy functions are compared. TI and DOS based evidence estimation both agree within the precision and give strong preference to the energy function that accounts for repulsive and attractive van der Waals contributions.

is $\log Z = 84.354$. Samples are generated at 11 equispaced inverse temperatures ranging from $\beta_{\min}$ to 1 where $\beta_{\min}$ is varied between 0 and the critical value β_c. For a given β, exact energy samples are drawn by using the analytical density of states calculated by Beale (1996). Figure 3(a) compares TI with DOS based evidence estimation for β_{min} approaching the critical value. As for the Gaussian likelihood (section 3.1), the DOS procedure is more stable if $\beta_{\min}$ deviates significantly from 0. For $\beta_{\min}$ approaching the critical value, of course, both methods become more and more inaccurate. TI tends to underestimate the log-evidence because positive contributions to the integral (5) are missing. The DOS method tends to overestimate the log-evidence because the density of states with low likelihood is underestimated. This is evident from Figure 3(b) showing the true DOS and the estimated DOS obtained with increasingly large $\beta_{\min}$.

3.5 Model comparison in protein structure calculation

Proteins are complex systems that are challenging to simulate (Hansmann and Okamoto, 1999). We compared TI against DOS based evidence estimation for the small protein ubiquitin (for details see Habeck (2011)). The parameters θ are the 370 torsion angles that parametrize rotations about chemical bonds. The likelihood involves distances measured with nuclear magnetic resonance spectroscopy (Rieping *et al.*, 2005; Habeck *et al.*, 2005a). The prior distribution is the Boltzmann ensemble based on a physical energy function. Two different potential energy functions are compared. The first model involves an energy function that is purely repulsive and penalizes atom clashes (Linge and Nilges, 1999). The second model involves an energy function that has both repulsive and attractive non-bonded contributions (Kuhlman *et al.*, 2003).

We use a two-parameter parallel tempering scheme to sample from the posterior distribution defined over protein conformational space (Habeck *et al.*, 2005b). In addition to the annealed likelihood, the prior distribution is modified by using the Tsallis ensemble parametrized with parameter $q \geq 1$. The "inverse temperature" of the likelihood, β, is varied from 1 to 10^{-5} in the first 70 replicas at constant $q = 1$. In the last 30 replicas, q is increased from 1.0 to 1.06 at constant $\lambda = 10^{-5}$. Using this scheme the likelihood is gradually switched off in the first 70 replicas, in the remaining 30 replicas also the prior is switched off. To compare both energy functions, we calculate the log-evidence for both simulations using TI and the DOS

procedure. Table 3(c) lists the log-evidences obtained with TI and DOS estimation for both physical energy functions. The log-evidence estimates agree within the precision and give strong preference to the energy function that accounts for repulsive and attractive van der Waals contributions. This is physically reasonable and also consistent with the finding that the preferred energy function also results in more accurate protein structure (Habeck, 2011).

4 Connection to nested sampling

Nested sampling (NS) (Skilling, 2004) is a recent MCMC method that primarily aims to calculate the evidence and considers posterior samples as a by-product. The evidence is written as

$$Z = \int_0^1 L(X)\,\mathrm{d}X \approx \sum_i L_i\,(X_{i-1} - X_i) \qquad (22)$$

where $X(L)$ is the prior mass of states with likelihood above the contour L and directly related to the density of states:

$$\begin{aligned} X(e^{-E}) &= \int \Theta(L(\theta) - e^{-E})\,\pi(\theta)\,\mathrm{d}\theta \\ &= \int \Theta(E - E')\,g(E')\,\mathrm{d}E'. \end{aligned} \qquad (23)$$

That is, X is the cumulative distribution function of the DOS and representations (7) and (22) are identical. Nested sampling constructs a series of likelihood contours $L_i = e^{-E_i}$ by sampling n independent particles from the ensemble $\Theta(E_i - E(\theta))\,\pi(\theta)$. The particle in the highest energy state is stored as a sample and its energy is used to define the likelihood contour of the next iteration. The prior mass under each contour is estimated statistically to amount to $X_i = e^{-i/n}$. According to (22), each NS iteration contributes $L_i w_i$ with $w_i = X_{i-1} - X_i$ to the evidence.

4.1 Deceptive Gaussian mixture model

We now look at Neal's "deceptive" mixture model (Neal, 1996), a challenging sampling and evidence estimation task. This 2D mixture of Gaussians has a total of 4292 equally weighted components. Each component has variance $\sigma = 0.001$ in both directions which are uncorrelated. The mixture components are located at four centers $(\pm 15, \pm 15)$: 121 finely spaced components are located in the upper-right quadrant, 121 widely spaced components are in the upper-left quadrant, 2025 finely spaced components are in the lower-left quadrant, 2025 widely spaced components are in the lower-right quadrant. The likelihood function is given by

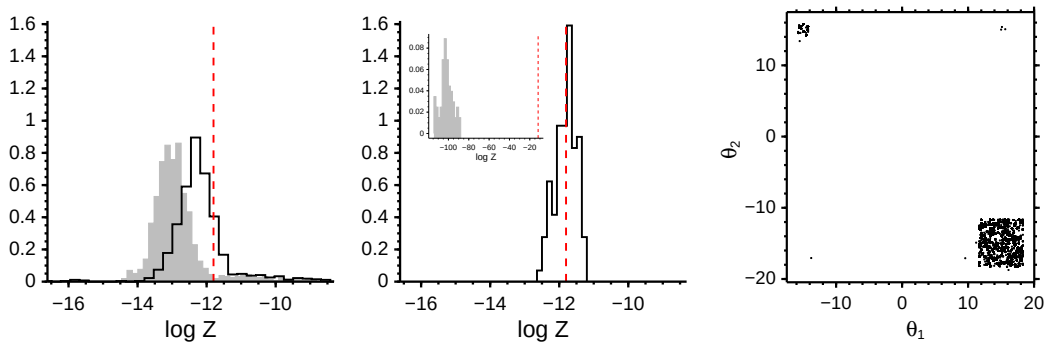
$$\begin{aligned} L(\theta) \;=\; & \sum_{i=-5}^{+5}\sum_{j=-5}^{+5} e^{-\frac{1}{2\sigma^2}\|\theta-\mu_{1,ij}\|^2} + \\ & \sum_{i=-5}^{+5}\sum_{j=-5}^{+5} e^{-\frac{1}{2\sigma^2}\|\theta-\mu_{2,ij}\|^2} + \\ & \sum_{i=-22}^{+22}\sum_{j=-22}^{+22} e^{-\frac{1}{2\sigma^2}\|\theta-\mu_{3,ij}\|^2} + \\ & \sum_{i=-22}^{+22}\sum_{j=-22}^{+22} e^{-\frac{1}{2\sigma^2}\|\theta-\mu_{4,ij}\|^2} \end{aligned} \qquad (24)$$

where $\mu_{1,ij} = (i\delta + 15, j\delta + 15)^T$, $\mu_{2,ij} = (i\Delta - 15, j\Delta + 15)^T$, $\mu_{3,ij} = (i\delta - 15, j\delta - 15)^T$ and $\mu_{4,ij} = (i\Delta + 15, j\Delta - 15)^T$ with a fine spacing $\delta = 2.5 \times 10^{-3}$ and a wide spacing $\Delta = 1.5 \times 10^{-1}$ between the components. The prior domain is bounded to $[-30, +30]$ in both dimensions. The evidence can be calculated analytically: $\log Z = \log(8584\,\pi\sigma^2) - 2\log(60) = -11.8$.

First, we run nested sampling with $n = 100$ particles and stop at the iteration whose relative contribution to the evidence ($L_i w_i / \sum_{j \geq i} L_j w_j$) is less than 10^{-5} (the resulting average number of iterations is ~ 2000). We sample states under a likelihood constraint by using a random walk in 2D confined to the support of the prior. States with likelihood below the current contour are rejected. An adaptive stepsize controls the range of the random walk. The state with highest energy is stored as posterior sample and replaced with a particle that is selected randomly from the remaining $n-1$ particles. The selected particle is then perturbed using 10 random walk steps. We can also analyze the samples obtained during nested sampling with our DOS estimation method. To this end, we store, in addition to the highest energy sample, the new state obtained after local perturbation.

Figure 4(a) shows a histogram of evidence estimates obtained with nested sampling and by analyzing the NS samples using our DOS procedure. Nested sampling's evidence estimate is biased toward values that are smaller than the true evidence, which may be the result of stopping the NS iterations too early. The DOS analysis, on the other hand, is systematically closer to the true evidence at the expense of producing more outliers (i.e. bins which are very distant from the bulk of the distribution).

For comparison, we also run a parallel tempering simulation with 10 temperatures using the annealing scheme of section 3.2 with exponent $c = 8$. Figure 4(b) shows the evidence estimates obtained by parallel tempering using thermodynamic integration and the density of states. Although the annealing sched-



(a) (b) (c)

Figure 4: Comparison of evidence estimates for the deceptive Gaussian mixture model. (a): Nested sampling results. Histograms are calculated over 1000 repetitions of nested sampling with 100 particles. The log-evidence obtained with nested sampling is shown as grey filled histogram, the estimate obtained with our DOS procedure is shown as black curve; the true evidence is marked with a dashed red line. (b): Parallel tempering results. 20000 parallel tempering exchanges were simulated, after a burnin of 100 exchanges the simulation has converged. At each temperature, every 100th sample after burnin is pooled into a single data set, resulting in 100 different data sets, whose evidence estimates are shown as histograms. The evidence estimate obtained with our DOS procedure is shown as black curve; the true evidence is marked with a dashed red line. The inset displays the evidence estimates obtained with thermodynamic integration. (c): Posterior samples obtained with parallel nested sampling.

ule was not optimized and produced very inhomogeneous exchange rates ranging from 5% to 68%, the DOS based evidence estimate achieves a better accuracy than nested sampling. For this simulation, thermodynamic integration fails spectacularly (see inset of Fig. 4(b)) mainly because contributions at small inverse temperatures dominate the evidence integral (5). As we increase the number of temperatures, thermodynamic integration eventually becomes as accurate as nested sampling or DOS based evidence estimation.

5 Conclusions

Marginal likelihood estimation via the density of states is an efficient and versatile alternative to thermodynamic integration, over which it has several advantages. When working with annealed posteriors, DOS based evidence estimation is more robust and flexible in the choice of the annealing schedule than thermodynamic integration. Moreover, we are not restricted to use samples from the actual posterior only, knowledge of the density of states allows us to combine samples from posteriors at all temperatures. It is straightforward to show that to calculate other marginal distributions, we simply need to expand the energy histogram with samples of the quantity of interest. DOS based evidence estimation also works with ensembles that do not bridge between the prior and the posterior, in which case thermodynamic integration cannot be applied at all. An example of such a family of distributions are the likelihood-bounded priors constructed during nested sampling. Similar to the method proposed here, nested sampling estimates the evidence via the density of states. To combine nested sampling with the proposed DOS estimation algorithm will be the subject of future research.

Acknowledgements

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) grant HA 5918/1-1 and the Max Planck Society.

References

Beale, P. D. (1996). Exact Distribution of Energies in the Two-Dimensional Ising Model. *Phys. Rev. Lett.*, **76**, 78–81.

Ferrenberg, A. M. and Swendsen, R. H. (1989). Optimized Monte Carlo Data Analysis. *Phys. Rev. Lett.*, **63**, 1195–1198.

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

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163–185.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statis-*

tics: Proceedings of the 23rd Symposium on the Interface, pages 156–163.

Habeck, M. (2011). Statistical mechanics analysis of sparse data. *J. Struct. Biol.*, **173**, 541–548.

Habeck, M., Nilges, M., and Rieping, W. (2005a). Bayesian inference applied to macromolecular structure determination. *Phys. Rev. E*, **72**, 031912.

Habeck, M., Nilges, M., and Rieping, W. (2005b). Replica-Exchange Monte Carlo scheme for Bayesian data analysis. *Phys. Rev. Lett.*, **94**, 0181051–0181054.

Hansmann, U. H. E. and Okamoto, Y. (1999). New Monte Carlo algorithms for protein folding. *Curr. Opin. Struct. Biol.*, **9**, 177–183.

Jeffreys, H. (1939). *Theory of Probability.* Oxford University Press.

Kass, R. and Raftery, A. (1995). Bayes factors. *American Statistical Association*, **90**, 773–775.

Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, **3**, 300–313.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.

Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A., and Rosenberg, J. M. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. *J. Comp. Chem.*, **13**, 1011–1021.

Li, Y., Protopopescu, V. A., Arnold, N., Zhang, X., and Gorin, A. (2009). Hybrid parallel tempering and simulated annealing method. *Applied Mathematics and Computation*, **212**(1), 216 – 228.

Linge, J. P. and Nilges, M. (1999). Influence of nonbonded parameters on the quality of NMR structures: a new force-field for NMR structure calculation. *J. Biomol. NMR*, **13**, 51–59.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge UK.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, **6**, 353–366.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, **11**, 125–139.

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential Structure Determination. *Science*, **309**, 303–306.

Shirts, M. R. and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys*, **129**, 124105.

Sivia, D. and Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial*. Oxford University Press, USA, 2nd edition.

Skilling, J. (2004). Nested sampling. *AIP Conference Proceedings*, **735**(1), 395–405.

Swendsen, R. H. and Wang, J.-S. (1986). Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.*, **57**, 2607–2609.