Prior-predictive value from fast-growth simulations: Error analysis and bias estimation

Alberto Favaro,* Daniel Nickelsen, Elena Barykina, and Andreas Engel
Institut für Physik, Carl-von-Ossietzky-Universtität, 26111 Oldenburg, Germany
(Received 3 June 2014; published 15 January 2015)

Variants of fluctuation theorems recently discovered in the statistical mechanics of nonequilibrium processes may be used for the efficient determination of high-dimensional integrals as typically occurring in Bayesian data analysis. In particular for multimodal distributions, Monte Carlo procedures not relying on perfect equilibration are advantageous. We provide a comprehensive statistical error analysis for the determination of the prior-predictive value (the evidence) in a Bayes problem, building on a variant of the Jarzynski equation. Special care is devoted to the characterization of the bias intrinsic to the method and statistical errors arising from exponential averages. We also discuss the determination of averages over multimodal posterior distributions with the help of a consequence of the Crooks relation. All our findings are verified by extensive numerical simulations of two model systems with bimodal likelihoods.

DOI: 10.1103/PhysRevE.91.012127 PACS number(s): 02.50.−r, 05.10.−a, 02.70.Uu

I. INTRODUCTION

Statistical data analysis is at the heart of all quantitative science. Observations, measurements, and numerical simulations alike are prone to random perturbations, and effort and care are needed to scrutinize the influence of these noisy disturbances on the results of the respective investigation. A particularly clear and efficient procedure to do so is provided by Bayesian inference [1–3]. In a typical setup, a model $\mathfrak{M}$ specified by parameters x is checked against observational, experimental, or numerical data d. All information on the parameters already available from previous experience is subsumed in the *prior* distribution $p_p(x|\mathfrak{M})$ of the parameters. The model itself is characterized by a *likelihood* distribution $p_l(d|x,\mathfrak{M})$ specifying the probability of data conditioned on a particular choice of the parameters. The application of Bayes rule,

$$p_{\text{post}}(x|d,\mathfrak{M}) = \frac{p_p(x|\mathfrak{M})p_l(d|x,\mathfrak{M})}{p(d|\mathfrak{M})}, \tag{1}$$

then yields the *posterior* distribution $p_{\text{post}}(x|d,\mathfrak{M})$ for the parameters x. It provides the statistically optimal combination of the information about the parameters contained in the prior and in the new data. Bayesian methods are being used for various problems in quite diverse fields of research [4–7]. They are particularly appropriate for testing null hypotheses [8] and in problems of model selection [2].

A crucial problem in concrete applications of Bayesian inference is the determination of the denominator in (1),

$$p(d|\mathfrak{M}) := \int p_p(x|\mathfrak{M})p_l(d|x,\mathfrak{M})\,\mathrm{d}^n x, \tag{2}$$

which is called *evidence*, or *prior-predictive value*, or *marginal likelihood*. Typically, the integral extends over a high-dimensional parameter space, and it is dominated by contributions from small and labyrinthine regions. This makes straight Monte Carlo methods rather inefficient [9].

Moreover, depending on the random data $d = \{d_i\}$, the evidence is itself a random quantity. As a rule, the likelihood is of the form

$$p_l(d|x,\mathfrak{M}) \sim \exp\left(\sum_i f(d_i)\right) \tag{3}$$

with some function $f(d)$, i.e., it factorizes in the individual data d_i. Since these data are commonly assumed to be independent, the evidence (2) is roughly a product of many independent random terms. As such, its probability distribution is similar to a log-normal distribution with a long tail implying a large difference between the average and the most probable value. Consequently, the whole distribution of the evidence as well as a typical value are badly characterized by the average.

Contrary, the log-evidence, $\ln p(d|\mathfrak{M})$, has the structure of a *sum* of independent random terms, and therefore, for large data sets, its distribution may be expected to be similar to a Gaussian one. Then, the most probable value and the typical value are similar to each other, and the distribution is well characterized by its first cumulants. This is the reason why the statistical characteristics of the log-evidence rather than those of the evidence itself are of central importance in Bayesian inference. Considering $\ln p(d|\mathfrak{M})$, as opposed to $p(d|\mathfrak{M})$, also makes it easier to draw links with the studies [10,11], which have guided our analysis.

Similar problems arise in statistical mechanics in the determination of partition functions as compared to free energies; see Sec. III A of [12]. It is hence not surprising that methods developed in statistical physics are being increasingly used in data analysis. A prominent example is thermodynamic integration [13], which is meanwhile routinely implemented in Bayesian inference [7,9,14]. Its applicability rests on the accurate determination of thermal averages of the logarithm of the likelihood distribution. This is a standard problem in computational physics and can often be accomplished by Markov chain Monte Carlo methods [7,15]. Nevertheless, for *multimodal* distributions, the relaxation times to thermal equilibrium can be very large, which may compromise the determination of the necessary averages. In fact, for a model system with a bimodal likelihood distribution, thermodynamic integration was shown to have substantial difficulties in determining the evidence $p(d|\mathfrak{M})$ of a Bayes problem [16].

*alberto.favaro@uni-oldenburg.de

There are several situations in which multimodal distributions occur quite naturally in Bayesian inference. A well-documented case is the determination of the relative phase between two interferometers in the presence of noise [17]. Plotting the two sinusoidal signals against each other results in an ellipse, the ellipticity of which determines the relative phase. Given the additional constraints present, there remain *two* possible ellipses for each data point; the corresponding likelihood distribution is hence bimodal. More complex situations are *mixture* models, which allow for an arbitrary number of components [18]. Problems of Monte Carlo methods for such mixture models are discussed, e.g., in [19].

In recent years, there have been fascinating developments in the statistical mechanics of nonequilibrium systems that gave rise to the emerging field of stochastic thermodynamics [20–22]. Central to this field are so-called work and fluctuation theorems, which, among other things, may be used to determine free-energy differences from nonequilibrium trajectories [23–25]. Because of the close relation between free-energy estimates and calculations of the log-evidence $\ln p(d|\mathfrak{M})$, these developments also bring about new possibilities for Bayesian data analysis [16]. In an inference problem, the nonequilibrium aspect is exhibited by the use of *nonstationary*, explicitly time-dependent Markov processes that do not rely on repeated equilibrations. Accordingly, when multimodal distributions are considered, these methods can prove advantageous.

In [16], a method was proposed to determine the evidence in a Bayes problem by using a variant of the Jarzynski equation [26,27]. In the present paper, we provide a detailed error analysis of this method when used to estimate the log-evidence in a Bayesian analysis. Due to the nonlinearities involved, the method has a bias and resilient statistical errors [11,28,29] which need to be treated with care. We also detail the calculation of averages over multimodal posteriors using a consequence of the Crooks relation [30].

The paper is organized as follows. In Sec. II, we provide the basic equations and fix the notation. In Sec. III, we present a detailed error analysis of the method for determining the log-evidence. Section IV demonstrates the performance of the proposed error analysis by means of two examples: a bimodal likelihood distribution composed of two Gaussians [9], and a similar likelihood distribution but composed of two Lorentzians [31]. Section V provides an analogous analysis for averages with the posterior distribution. Finally, Sec. VI contains our conclusions.

II. BASIC EQUATIONS

In the following, the dependence of the prior and the likelihood distribution on the parameters x of the model is the important one. We therefore temporarily suppress the dependence on d and $\mathfrak{M}$ for notational convenience.

For a successful application of Bayesian inference in problems of practical relevance, effective numerical methods are crucial. It is well known that normalization factors of distributions, such as the evidence $p(d|\mathfrak{M})$, are much harder to obtain using Monte Carlo methods than the corresponding averages [15]. It is therefore desirable to replace the integration in (2) by functions of such averages. A simple method to do so is the following variant of *thermodynamic integration* [13].

Defining the auxiliary quantity

$$Z(\beta) := \int [p_l(x)]^\beta \, p_p(x) \, \mathrm{d}^n x, \tag{4}$$

we have $Z(0) = 1$ due to the normalization of the prior distribution and $Z(1) = p(d|\mathfrak{M})$, which is the desired evidence. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \ln Z(\beta) = \frac{1}{Z(\beta)} \int \ln p_l(x) \; p_l^\beta(x) \, p_p(x) \, \mathrm{d}^n x. \tag{5}$$

The right-hand side of this equation denotes the average $\langle \ln p_l(x) \rangle_\beta$ of the log-likelihood distribution with

$$P_\beta(x) := \frac{1}{Z(\beta)} \; p_l^\beta(x) \, p_p(x). \tag{6}$$

Hence,

$$\begin{aligned} \ln p(d|\mathfrak{M}) &= \int_0^1 \mathrm{d}\beta \, \frac{\mathrm{d}}{\mathrm{d}\beta} \ln Z(\beta) \\ &= \int_0^1 \mathrm{d}\beta \, \langle \ln p_l(x) \rangle_\beta . \end{aligned} \tag{7}$$

In practical applications of this relation, one chooses $n = 10, \ldots, 100$ values β_n from the interval $(0,1)$ and calculates the averages $\langle \ln p_l(x) \rangle_{\beta_n}$ by standard Markov chain Monte Carlo (MCMC) sampling. The implemented transition probability $\rho(x,x';\beta_n)$ of the Markov chain has to be consistent with the corresponding stationary distribution (6). This is most directly ensured by the detailed balance condition [15]. Having obtained the n averages $\langle \ln p_l(x) \rangle_{\beta_n}$, the integral in (7) can be determined approximately. We note that the Markov chain used for each of the β values is stationary, i.e., there is no explicit time dependence in the transition probability $\rho(x,x';\beta_n)$.

This variant of thermodynamic integration works fine as long as there are no difficulties with the equilibration of the individual Monte Carlo runs [9]. However, for multimodal distributions, problems may arise due to trajectories getting stuck in local maxima of the distribution [19]. In the generic case of unimodal prior and multimodal likelihood distributions, such problems show up when β approaches 1. The last points for the calculation of the integral in (7) are then prone to errors, and the estimate for the evidence $p(d|\mathfrak{M})$ becomes unreliable.

These equilibration problems may be circumvented by building on modern methods for free-energy estimation that use nonstationary trajectories [25,26]. Toward that end, one considers a finite time interval $t \in (0,T)$ in which β changes from 0 to 1. In the numerics, this is done by fixing a set of intermediate times and corresponding increments $\{t_m, \Delta\beta_m\}$, the so-called protocol $\beta(t)$. Starting from a point x_0 sampled from the prior distribution, MCMC simulations with the time-dependent transition rate $\rho(x,x';\beta(t))$ are performed. For each realization $x(t)$ of such a simulation, one determines the quantity

$$R[x(\cdots)] = \sum_m \Delta\beta_m \; \ln p_l[x(t_m)]. \tag{8}$$

As shown in [16], one then finds the exact relation

$$\langle e^R \rangle = Z(1) = p(d|\mathfrak{M}), \tag{9}$$

where the average in (9) is over independent realizations $x(t)$ of the nonstationary Markov process. In nonequilibrium thermodynamics, the above relation is known as the *Jarzynski equation*. The continuum version of (8) has the form (see also [32–34])

$$R[x(\cdots)] = \int_0^T \frac{\partial}{\partial t}\,\beta(t)\,\ln p_I[x(t)]\,\mathrm{d}t. \tag{10}$$

Notably, averages with the posterior distribution may be expressed in a similar way. For a reasonable function $f(x)$, one can show that [16]

$$\langle f\rangle_{\mathrm{post}} = \int f(x)\,p_{\mathrm{post}}(x)\,\mathrm{d}x = \frac{\langle e^R f[x(T)]\rangle}{\langle e^R\rangle}, \tag{11}$$

where $x(T)$ denotes the final point of the trajectory $x(t)$, and the averages are again over an ensemble of realizations. We remark that (11) is a consequence of the Crooks relation; see Eq. (21) in [30].

III. ERROR ANALYSIS OF THE JARZYNSKI ESTIMATOR

The Jarzynski equation (9) to determine the evidence from nonstationary realizations $x(t)$ involves the exponential average

$$\langle e^R\rangle := \int \mathrm{d}x\; p(R)\,e^R. \tag{12}$$

In practice, the distribution $p(R)$ of the random variable R is unknown, and $\langle\cdots\rangle$ is replaced by an ensemble average,

$$\langle e^R\rangle_M := \frac{1}{M}\sum_{i=1}^{M} e^{R_i}, \tag{13}$$

where the index M in $\langle e^R\rangle_M$ denotes the number of samples R_i that contribute to $\langle e^R\rangle_M$.

Replacing the exact average (12) with the sample mean (13) introduces an error that vanishes in the limit of infinitely many samples, $M\to\infty$. However, due to the exponential weight on large R values invoked by the nonlinear average, this error may remain significant even for large M. In addition, estimating the logarithm of the evidence generates a bias in the statistics of $\ln\langle e^R\rangle_M$ [10,11,25].

The analysis of these errors is the central subject of this paper and will be discussed in this section. The first part concerns the bias of the random variable $\ln\langle e^R\rangle_M$ on the basis of exact averages. The second part includes the error for considering finite-sized ensembles of e^R.

A. Basic notions

To compute the log-evidence from an M-sized ensemble of R values, we use (9) and (13) to define the *Jarzynski estimator*,

$$\ln p(d|\mathfrak{M}) \simeq \ln\langle e^R\rangle_M = \ln\frac{1}{M}\sum_{i=1}^{M} e^{R_i}. \tag{14}$$

Considering several M-sized ensembles of R values, the sample mean $\langle e^R\rangle_M$ is a random variable for any finite M. The statistics of $\ln\langle e^R\rangle_M$ is central to our error analysis of the Jarzynski estimator. To assess the statistics of $\ln\langle e^R\rangle_M$, we define bias B, variance σ^2, and mean square error α^2 as

$$B(M) := \langle\ln\langle e^R\rangle_M\rangle - \ln p(d|\mathfrak{M}), \tag{15}$$

$$\sigma^2(M) := \langle(\ln\langle e^R\rangle_M - \langle\ln\langle e^R\rangle_M\rangle)^2\rangle, \tag{16}$$

$$\alpha^2(M) := \langle[\ln\langle e^R\rangle_M - \ln p(d|\mathfrak{M})]^2\rangle. \tag{17}$$

It is worth noting that these quantities are related by

$$\alpha^2(M) = \sigma^2(M) + B^2(M). \tag{18}$$

To understand why a nonzero bias (15) may occur, a valid starting point is

$$\langle\langle e^R\rangle_M\rangle = \langle e^R\rangle = p(d|\mathfrak{M}). \tag{19}$$

One substitutes this identity into the definition for the bias (15), and establishes that

$$B(M) = \langle\ln\langle e^R\rangle_M\rangle - \ln\langle\langle e^R\rangle_M\rangle. \tag{20}$$

Hence, a finite bias signals that the logarithm and the expectation value do not commute.

A related statement can be derived from the Jensen inequality [35]. If the function φ is convex on the interval I, and X is a stochastic variable with range $J\subseteq I$, then

$$\langle\varphi(X)\rangle \geqslant \varphi(\langle X\rangle). \tag{21}$$

When $\varphi(X) = -\ln(X)$ and $X = \langle e^R\rangle_M$, the inequality (21) prescribes that

$$\langle\ln\langle e^R\rangle_M\rangle \leqslant \ln\langle\langle e^R\rangle_M\rangle. \tag{22}$$

Thus, according to (20), the bias of the Jarzynski estimator is negative, or zero. For the analogous property in statistical physics, we refer to [10,36].

B. Confidence interval

If the bias B and the variance σ^2 as defined in (15) and (16) are known, the root mean square error α follows from (18) and serves as a measure of uncertainty for the estimation of the log-evidence, $\ln p(d|\mathfrak{M})$. While the computation of σ^2 from finite samples is straightforward, the determination of B is intricate as it involves $p(d|\mathfrak{M})$ itself. It therefore is common practice to substitute $p(d|\mathfrak{M})$ with an appropriate estimator, in the case at hand being $p(d|\mathfrak{M})\simeq\langle e^R\rangle_M$. The consequence is that the resulting α only accounts for the bias generated by the logarithm in the Jarzynski estimator (14), and not for the nonlinearity of the exponential average. In what follows, the full bias B will be split into two contributions C and D, in which C uses the mentioned substitution $p(d|\mathfrak{M})\simeq\langle e^R\rangle_M$, and D takes care of the error brought about by this step.

In tackling the intrinsic problem that the true values of $p(d|\mathfrak{M})$, and therefore also B, are not known, the key point will be to derive a confidence interval for D from the central limit theorem [35]. To do so, we make two assumptions:

(i) $\{e^{R_1},\ldots,e^{R_N}\}$ is a sequence of N independent random variables that have the same distribution.

(ii) The variance ς^2 of that distribution is finite—while the expectation value is $p(d|\mathfrak{M})$, because of (9).

The sample size N, in addition to M, is introduced for later convenience, and we assume that $N\gg M$. Note that (ii) refers

to the distribution of e^R, the variance of which may be finite despite likelihood distributions with infinite variance. We will demonstrate this point in Sec. IV B.

If (i) and (ii) are satisfied, the central limit theorem dictates that, as N approaches infinity, the random variable

$$Y(N) := \sqrt{N}[\langle e^R\rangle_N - p(d|\mathfrak{M})]/\varsigma \tag{23}$$

becomes normally distributed, with zero mean and unit variance. Accordingly, a confidence interval for $Y(N)$ may be written as

$$\Pr[-\sqrt{2}\,\mathrm{erf}^{-1}(\gamma) < Y(N) < \sqrt{2}\,\mathrm{erf}^{-1}(\gamma)] \approx \gamma, \tag{24}$$

where $\Pr[\cdots]$ indicates probability, and erf^{-1} is the inverse error function. The confidence level γ can be selected as one deems fit, but ordinary choices are 95%, 99%, 99.5%, and 99.9%; see [37]. Throughout this paper, we will use the rather pessimistic choice $\gamma = 0.95$. The approximate sign in (24) accounts for the fact that N is taken to be finite.

The confidence interval for $Y(N)$ can be transferred to the bias B. Toward that end, we solve (23) for $p(d|\mathfrak{M})$, and we substitute the result into (15). Hence, the bias is expressed as

$$B(M) = C(M,N) + D(N), \tag{25}$$

with

$$C(M,N) = \langle \ln\langle e^R\rangle_M\rangle - \ln\langle e^R\rangle_N, \tag{26}$$

$$D(N) = -\ln\left[1 - \frac{\varsigma}{\sqrt{N}}\frac{Y(N)}{\langle e^R\rangle_N}\right]. \tag{27}$$

The dependency on N of the first term in (25) is compensated by the second term. We multiply the inequality within brackets in (24) by the positive quantity $\varsigma/\sqrt{N}\langle e^R\rangle_N$. Then, to incorporate $D(N)$, we apply the monotonic increasing function $-\ln(1-X)$; see (27). Both these operations do not reverse the sign of the inequality. It follows that

$$\Pr[D_-(N) < D(N) < D_+(N)] \approx \gamma, \tag{28}$$

with

$$D_\pm(N) = -\ln\left[1 \mp \sqrt{\frac{2}{N}}\frac{\varsigma\,\mathrm{erf}^{-1}(\gamma)}{\langle e^R\rangle_N}\right]. \tag{29}$$

Finally, by adding $C(M,N)$ to the inequality within brackets in (28), a confidence interval for the bias is attained,

$$\Pr[B_-(M,N) < B(N) < B_+(M,N)] \approx \gamma, \tag{30}$$

with the confidence limits

$$B_\pm(M,N) = C(M,N) + D_\pm(N). \tag{31}$$

Two comments are in order. First, when N is large enough, one has that

$$0 < \sqrt{\frac{2}{N}}\frac{\varsigma}{\langle e^R\rangle_N} < 1. \tag{32}$$

The inequality on the left always holds true. Because γ is positive, $\mathrm{erf}^{-1}(\gamma)$ ranges from 0 to 1, and

$$0 < \sqrt{\frac{2}{N}}\frac{\varsigma\,\mathrm{erf}^{-1}(\gamma)}{\langle e^R\rangle_N} < 1. \tag{33}$$

This ensures that the confidence limit $D_+(N)$ is finite and real; cf. (29). Second, in (31), the dependency of $C(M,N)$ on N is not compensated by that of $D_\pm(N)$. It follows that the confidence limits $B_\pm(M,N)$ are functions of M and also N.

We are now in the position to derive a confidence interval for the mean-square error $\alpha^2(M)$; see (18). Motivated by the procedure followed earlier on, it is natural to define

$$\alpha^2_\pm(M,N) = \sigma^2(M) + B^2_\pm(M,N). \tag{34}$$

Unfortunately, this is not a monotonic function, and the direction of previous inequalities gets mixed up. Nevertheless, it is still possible to conclude that

$$\Pr[\alpha^2(M) < \max[\alpha^2_+(M,N),\alpha^2_-(M,N)]] \gtrsim \gamma, \tag{35}$$

where $\max[\cdots]$ selects the larger of its two arguments.

The error analysis proposed above involves the exact averages $\langle\cdots\rangle$. For practical purposes, however, it is necessary to estimate the averages $\langle\cdots\rangle$ by empirical averages as defined in (13). To do so, we take N as the given total number of R values, group these into N/M blocks of size M, and estimate

$$\langle\langle\cdots\rangle_M\rangle \simeq \langle\langle\cdots\rangle_M\rangle_{\frac{N}{M}}. \tag{36}$$

This procedure, commonly referred to as block-averaging, was pioneered by Wood, Mühlbauer, and Thompson [38]. We mention that an alternative to block-averaging is the bootstrap algorithm, as explored in the article [29] by Ytreberg and Zuckerman.

In the remaining part of the paper, we will use the prescription (36) to estimate $C(M,N)$ and $\sigma^2(M)$, defined in (26) and (16), from simulation results of an N-sized ensemble of R values. To estimate $D(N)$ and $D_\pm(N)$, defined in (27) and (29), as well as the confidence interval for the bias in (30) and the mean square error in (35), we approximate the variance ς^2 of the distribution for e^R with the sample variance

$$\hat{\varsigma}^2(N) := \frac{1}{N-1}\sum_{i=1}^{N}(e^{R_i} - \langle e^R\rangle_N)^2. \tag{37}$$

Likewise, for σ^2, we take

$$\hat{\sigma}^2(M,N) := \frac{1}{N/M-1}\sum_{i=1}^{N/M}\left[\ln\langle e^R\rangle_M - \langle\ln\langle e^R\rangle_M\rangle_{\frac{N}{M}}\right]^2. \tag{38}$$

We will denote estimated quantities that use block-averages and sample variances instead of exact averages with a "hat," for instance,

$$\hat{B}(M,N) = \langle\ln\langle e^R\rangle_M\rangle_{\frac{N}{M}} - \ln p(d|\mathfrak{M}), \tag{39}$$

$$\hat{C}(M,N) = \langle\ln\langle e^R\rangle_M\rangle_{\frac{N}{M}} - \ln\langle e^R\rangle_N, \tag{40}$$

$$\hat{D}_\pm(M,N) = -\ln\left[1 \mp \sqrt{\frac{2}{N}}\frac{\hat{\varsigma}(N)\,\mathrm{erf}^{-1}(\gamma)}{\langle e^R\rangle_N}\right], \tag{41}$$

$$\hat{\alpha}^2(M,N) = \hat{\sigma}^2(M,N) + \hat{B}^2(M,N), \tag{42}$$

in contrast to the exact expressions (27), (26), (29), and (18). The confidence limits $\hat{D}_\pm(M,N)$, as opposed to $D(N)$, are independent of the unknown $p(d|\mathfrak{M})$. Accordingly, the same holds true for

$$\hat{\alpha}^2_\pm(M,N) = \hat{\sigma}^2(M,N) + [\hat{C}(M,N) + \hat{D}_\pm(M,N)]^2. \tag{43}$$

IV. EXAMPLES FOR THE ESTIMATION OF THE LOG-EVIDENCE

Section III was devoted to the bias B of the Jarzynski estimator (14). We split the bias into two components, $B = C + D$, where C is treated by block-averaging and D is the remaining unknown discrepancy of the estimator. Based on the central limit theorem, we derived the confidence limits $D_\pm$ for the unknown D.

In this section, we demonstrate the performance of the Jarzynski estimator and the proposed error analysis for two exactly solvable settings involving bimodal likelihood distributions. We also relate our error analysis to those existing in the literature, which exemplifies that C is useful to judge the applicability of the central limit theorem, indicating the minimum total number of R values for which $D_\pm$ becomes reliable.

A. Gaussian bimodal likelihood distribution

To construct a bimodal $P_\beta(x)$, the simplest option appears to be that of setting the likelihood distribution $p_l(d|x,\mathfrak{M})$ to be the sum of two Gaussians [9]. Hence, we specify that

$$p_l(d|x,\mathfrak{M}) = q_1 G\big(x,d,\sigma_l^2\big) + q_2 G\big(x,-d,\sigma_l^2\big), \tag{44}$$

where x and d are vectors of dimension n, and q_1 and q_2 assign different weights to the Gaussians,

$$G(x,\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{(\mu-x)^2}{2\sigma^2}\right), \tag{45}$$

with mean vector μ and variance σ^2. Choosing values for q_1 and q_2 that differ substantially from each other makes the equilibration problem particularly pronounced: while the positions of the maxima become apparent rather quickly, sampling the maxima with the correct weights q_1 and q_2 is reliant on the very rare trajectories that cross the low-probability region between the maxima.

The benefit of the Gaussian model is that the evidence is known analytically—if the prior distribution is taken to be Gaussian. Notably, this choice for $p_p(x|\mathfrak{M})$ is widespread in the Bayesian inference literature. Thus, we demand that

$$p_p(x|\mathfrak{M}) = G\big(x,0,\sigma_p^2\big) \tag{46}$$

to find

$$p(d|\mathfrak{M}) = G\big(d,0,\sigma_p^2+\sigma_l^2\big). \tag{47}$$

We therefore have an analytic result that we can use to test our error analysis.

The dimension n will be set to 5, and d will be taken to have all of its components equal to 10. The maxima of the likelihood distribution are hence located at $\pm d = \pm(d_1,\ldots,d_5) = \pm(10,\ldots,10)$. For the weights q_i, we choose $q_1 = \frac{1}{21}$ and $q_2 = \frac{20}{21}$. The variances in (44) and (46) are selected to be $\sigma_l^2 = 1$ and $\sigma_p^2 = 100$, since in the typical Bayesian setup, the prior distribution is much broader than the likelihood distribution.

As discussed in Sec. II, the protocol β varies from 0 to 1 along every trajectory. We prescribe that β increases in a cubic way,

$$\beta = 0.05t + 0.95t^3, \tag{48}$$

where t is incremented from 0 to 1 in 25 steps. For each value of the protocol, the MCMC algorithm explores the parameter space, with 20 steps in the Markov chain. These values correspond to relatively short trajectories, whereby the computational resources can be focused on generating a large number N of R values.

In Ref. [28], Zuckerman and Woolf demonstrate that, when M is large,

$$-B(M) \approx \frac{\sigma^2(M)}{2} \approx \frac{1}{2M}\left[\frac{\varsigma}{p(d|\mathfrak{M})}\right]^2, \tag{49}$$

as a consequence of the central limit theorem. The same result is obtained in Ref. [11] by Gore, Ritort and Bustamante. It is worthwhile to observe that (49) involves only exact quantities. Accordingly, the above relation can be used to identify a threshold for M above which the central limit theorem for the random variable $\langle e^R\rangle_M$ may be applied. As the derivation of the confidence limits $D_\pm(N)$ rests on this very assumption, we conclude that $D_\pm(N)$ becomes reliable for values of N above the same threshold.

To identify for the introduced bimodal Gaussian example (44) the regime where the central limit theorem is applicable, we generated a total number of $N = 6\times10^7$ R values, and we substitute these in the numerically accessible variant of (49),

$$-\hat{C}(M) \approx \frac{\hat{\sigma}^2(M)}{2} \approx \frac{1}{2M}\left[\frac{\hat{\varsigma}(N)}{\langle e^R\rangle_N}\right]^2, \tag{50}$$

as used by Gore *et al.* in [11]. In Fig. 1, we plot the three quantities in (50) for all possible divisors M of $N = 6\times10^7$. The threshold above which the central limit theorem applies appears to be about $M \approx 10^4$; for $M > 10^4$, all quantities exhibit the predicted $1/M$ behavior. Therefore, our error analysis is in agreement with [11,28].

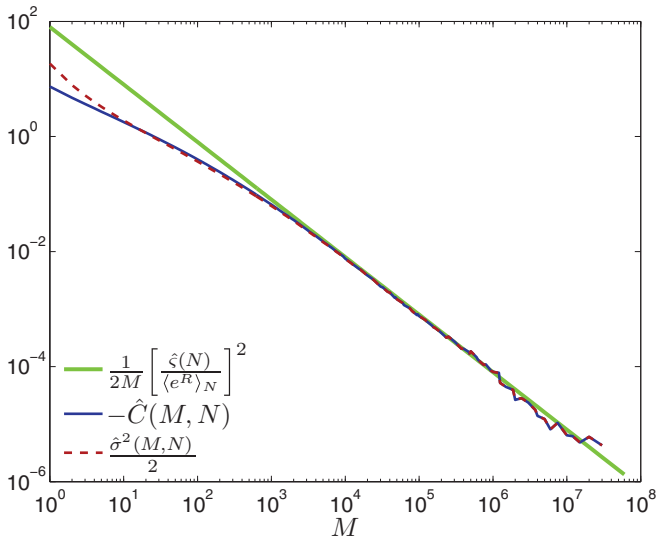

FIG. 1. (Color online) Verifying that our error analysis is consistent with the result (50) obtained in [11,28] based on the central limit theorem for the random variable e^R. The total number of Markov chains, and consequently of R values, is $N = 6\times10^7$.

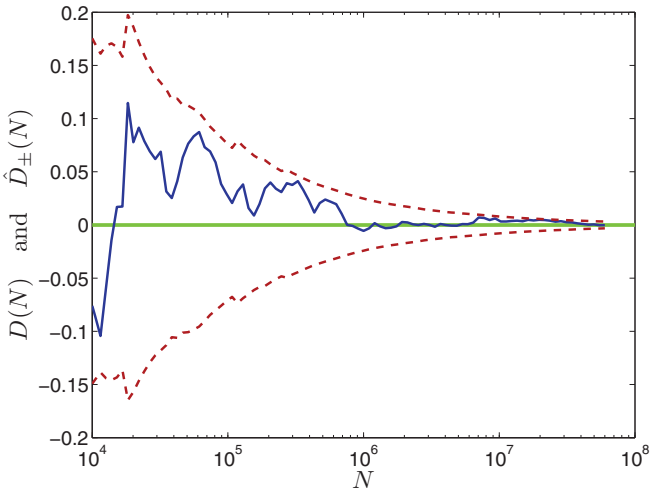

FIG. 2. (Color online) As the number of R values, N, gets larger, the confidence interval (28) gets smaller. Nevertheless, for the data set examined here, $D(N)$ stays within $\hat{D}_+(N)$ and $\hat{D}_-(N)$. The value of $D(N)$ follows from (27) using the exact value of $p(d|\mathfrak{M})$ given by (47); the estimated confidence limits $\hat{D}_\pm(N)$ are derived from (41).

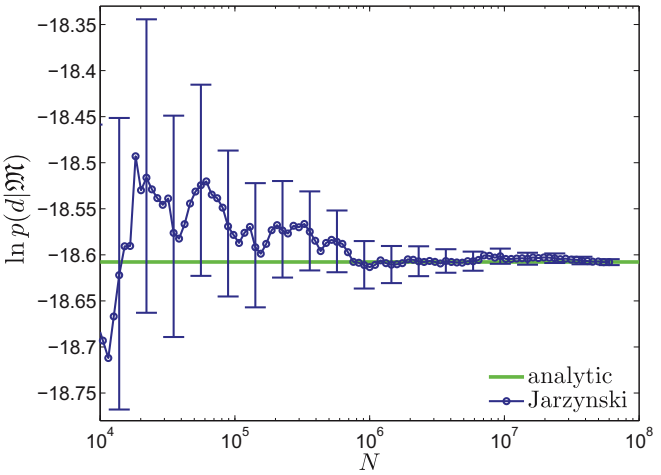

FIG. 3. (Color online) Estimation of the log-evidence for the bimodal Gaussian model (44) using the Jarzynski estimator (14) for an increasing number N of R values. The thick line is the analytic result (47), and the symbols use the Jarzynski estimator. The error bars are given by the estimates of the bias, $\hat{D}_\pm(N) = \hat{B}_\pm(N,N) = \hat{\alpha}_\pm(N,N)$, from (41).

To get an error margin for the estimation of $\ln p(d|\mathfrak{M})$, one could choose $M < N$ and use $\langle \ln\langle e^R\rangle_M\rangle_{N/M}$ as an estimator, for which $C(M,N)$ is an appropriate error measure. For the *single-block* estimator, that is, choosing $M = N$, we expect to be closest to the true value, but since $C(N,N) = 0$, the block-average procedure gives no result for the remaining bias $D(N)$. Instead, the confidence limits $D_\pm(N)$ may be used as an error margin for the single-block estimate. In Fig. 2, we plot $D(N)$ as well as $\hat{D}_\pm(N)$, estimated by using (41). The depicted range of N values is larger than the threshold 10^4 above which we assume the central limit theorem for $\langle e^R\rangle_N$ to hold and $D_\pm(N)$ to be reliable. Indeed, it is observed that $D_\pm(N)$ smoothly approach zero and that $D(N)$ belongs to the confidence interval (28).

Finally, in Fig. 3, we demonstrate the performance of the Jarzynski estimator and the proposed error analysis for an increasing number N of considered R values. For the single-block estimate of $\ln p(d|\mathfrak{M})$, i.e., $M = N$ and $\hat{\sigma}(N) = 0$, the estimated root mean square error is $\hat{\alpha} = \hat{B}$, and as furthermore $C(N,N) = 0$, it is simply $\hat{\alpha}_\pm = \hat{D}_\pm(N)$. For the smallest value of N, we again choose the threshold $N = 10^4$ above which the confidence limits $\hat{D}_\pm(N)$ are assumed to be reliable. We therefore use in Fig. 3 the limits $\hat{D}_\pm(N)$ as error bars, which are found to always cover the analytic result.

B. Likelihood distribution with infinite variance

The proposed error analysis in Sec. III B relies on the applicability of the central limit theorem to the random variable e^R, that is, a finite variance ς^2. As mentioned before, the requirement $\varsigma^2 < \infty$ can be satisfied by likelihood distributions with infinite variance. We demonstrate this in the present section.

Toward that end, we consider the Cauchy distribution (also known as a Lorentzian)

$$p(x) = \frac{s}{\pi}[s^2 + (x-d)^2]^{-1}. \tag{51}$$

Cauchy distributions are known to occur in power spectra of oscillating signals [31,39]. Helioseismic spectra to probe the interior of stars [40,41] are a recent example of a Bayesian analysis in which a multimodal likelihood distribution of the form (53) is used. For a limited number of data points, the posterior is typically multimodal itself due to peaks in the power spectra being artifacts of data processing or of instrumental origin.

The moments of the Cauchy distribution do not exist; in particular, the variance is divergent. Therefore, instead of a mean and a variance, the Cauchy distribution is characterized by the parameters d and s, where d is the mode of $p(x)$, and s specifies the width, as $2p(d+s) = p(d)$. The cumulative distribution is known analytically and reads

$$P(x) = \frac{1}{\pi}\arctan\left[\frac{x-d}{s}\right]. \tag{52}$$

To ensure a close analogy to the Gaussian example, we combine two Cauchy distributions to construct the bimodal likelihood distribution,

$$p_l(d|x) = \left(\frac{s}{\pi}\right)^n \left[q_1 \prod_{i=1}^{n}[s^2 + (d_i - x_i)^2]^{-1} + q_2 \prod_{i=1}^{n}[s^2 + (d_i + x_i)^2]^{-1}\right]. \tag{53}$$

Here, x is an n-dimensional parameter vector, and we take again one measurement d to be of the same dimension as x.

Similar to the Gaussian example, we choose the parameters $n = 5$, $d = (10,10,10,10,10)^{\mathrm{T}}$, $s = 0.1$, $q_1 = 20/21$, and $q_2 = 1/21$. To compute the evidence analytically from (52), we employ a flat prior on the interval $[-20s, 20s]$. The interval covers both modes of the likelihood distribution and therefore does not include any *a priori* information on the shape of the likelihood distribution; in fact, choosing a flat prior that does not cover the modes is found to drastically improve

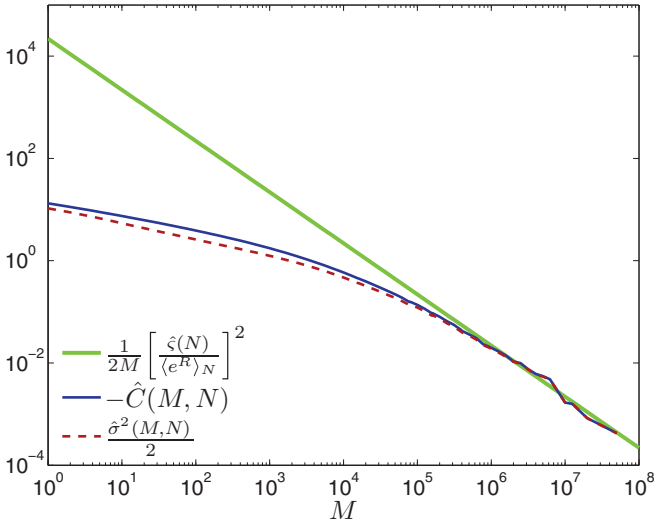

FIG. 4. (Color online) Verification for the likelihood distribution (53) involving two Cauchy distributions that our error analysis is consistent with the result (50) obtained in [11,28] based on the central limit theorem for the random variable e^R. The total number of Markov chains, and consequently of R values, is $N = 10^8$.

the performance of the Jarzynski method, since the Markov chains never start at one of the modes but instead run into the respective minima according to the weights q_1 and q_2. The protocol $\beta(t)$ is the same as for the Gaussian example; see (48).

We repeat the analysis of the Jarzynski estimator as done for the Gaussian example in the previous section, and we determine the log-evidence, $\ln p(d|\mathfrak{M})$, and error margins for the bimodal likelihood distribution defined in (53). To do so, we generate $N = 10^8$ Markov chains using again the MCMC algorithm, and we compute the corresponding R values from (8).

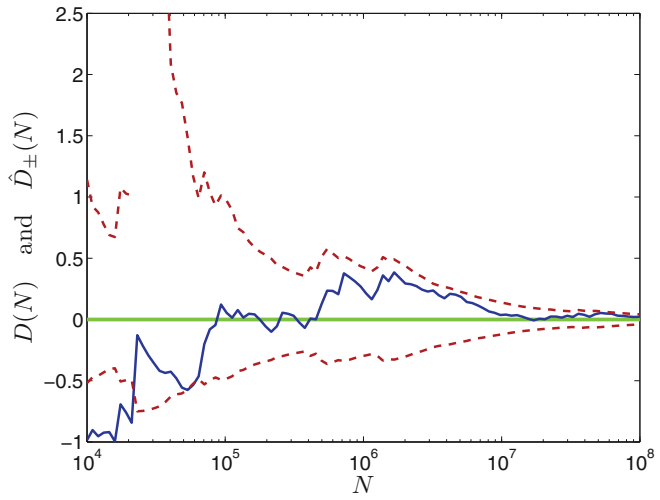

FIG. 5. (Color online) The value of $D(N)$ defined in (27) together with its confidence interval given by $\hat{D}_\pm(N)$ from (41) as a function of increasing number N of considered samples for R for the likelihood distribution in (53). The determination of $D(N)$ involves the exact value of $p(d|\mathfrak{M})$, which follows from using the cumulative Cauchy distribution in (52).

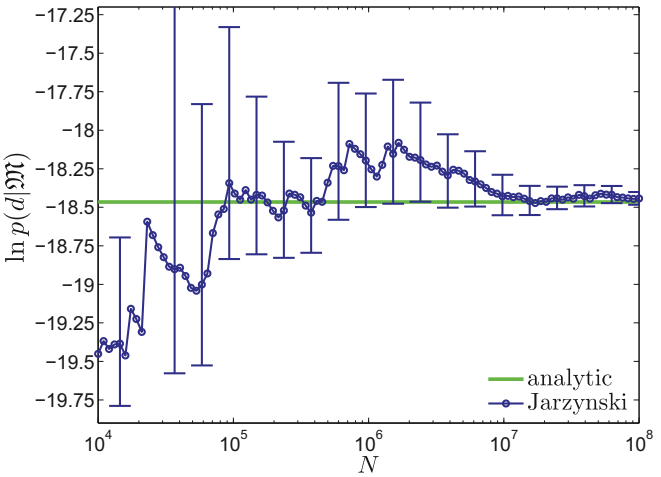

FIG. 6. (Color online) Estimation of the log-evidence for the likelihood distribution (53), which involves two Cauchy distributions, using the Jarzynski estimator (14) for an increasing number N of R values. The thick line is the analytic result obtained from (52); the symbols use the Jarzynski estimator. The error bars are given by the confidence limits $\hat{D}_\pm(N)$ from (41).

Figure 4 reveals that the central limit theorem holds for a number of more than about 10^6 Markov chains, cf. the discussion of the Gaussian example after Fig. 1 in the preceding subsection. We hence conclude that the variance of the random variable $\exp(R)$ is indeed finite.

The confidence limits of the bias of the single-block estimate for $\ln p(d|\mathfrak{M})$, being $\hat{D}_\pm(N)$ from (41), are depicted in Fig. 5 for an increasing number N of R values, together with $D(N)$ from (27) using the exact result of $p(d|\mathfrak{M})$ from (52). It is evident that for $N > 10^6$ the confidence limits $\hat{D}_\pm(N)$ smoothly approach zero enclosing $D(N)$.

Finally, in Fig. 6, we demonstrate the performance of the Jarzynski method and the proposed error analysis for increasing N. It is evident that $\hat{D}_\pm$ is again a well-suited error margin even for this example of a heavy tailed likelihood distribution, as the true value $\ln p(d|\mathfrak{M})$ is again always covered by $\hat{D}_\pm$.

V. AVERAGES WITH THE POSTERIOR DISTRIBUTION

We now focus on the problem of computing averages with respect to the posterior distribution numerically. Our aim is to investigate the fast-growth algorithm based on (11), which is closely related to the Jarzynski estimator (14). We demonstrate that the fast-growth calculations of $\langle\cdots\rangle_{\text{post}}$ are particularly advantageous when $p_{\text{post}}(x|d,\mathfrak{M})$ is multimodal. The severe problems that, under these circumstances, affect the standard Monte Carlo method are, to a large extent, overcome by the consequence (11) of the Crooks relation.

For the assessment, we make use of the bimodal Gaussian example described in Sec. IV A and consider the average of the function

$$f(x) = x_\parallel = \frac{x \cdot d}{|d|} \tag{54}$$

with respect to the posterior distribution. The scalar $x_\parallel$ is the component of the vector x along the vector d, specifying

the locations of the maxima in the posterior distribution. Our simulations are compared with the analytic result

$$\langle x_\parallel \rangle_{\text{post}} = (q_1 - q_2)\left(\frac{\sigma_p^2}{\sigma_p^2 + \sigma_l^2}\right)|d|, \tag{55}$$

which, for the parameter values used in Sec. IV A, gives

$$\langle x_\parallel \rangle_{\text{post}} \approx -20.0308. \tag{56}$$

To gain insight, it is useful to examine a standard Monte Carlo algorithm. Multiple stationary Markov chains are set to explore the parameter space, with $p_{\text{post}}(x|d,\mathfrak{M})$ as their target distribution. Along each trajectory, the average of $x_\parallel$ is calculated. Then, a further average across the Markov chains yields an estimate of $\langle x_\parallel \rangle_{\text{post}}$. In our simulation, 6×10^5 trajectories are generated, each with 5×10^4 steps. This makes a total of 3×10^{10} steps, and it corresponds to the estimate

$$\langle x_\parallel \rangle_{\text{post}} \approx -0.12. \tag{57}$$

It is evident that the standard Monte Carlo algorithm fails to solve the problem at hand. The histogram in Fig. 7 explains the reason for such a failure. Although the two peaks of the posterior distribution have different weights, $q_1 = \frac{1}{21}$ and $q_2 = \frac{20}{21}$, they contribute to (57) roughly in equal measure. More specifically, one can determine that the chains get trapped around the maxima of $p_{\text{post}}(x|d,\mathfrak{M})$.

Let us investigate the fast-growth estimator

$$\langle x_\parallel \rangle_{\text{post}} \approx \frac{\langle e^R x_\parallel(T)\rangle_N}{\langle e^R \rangle_N}, \tag{58}$$

see (11) and (54). As before, $\langle \cdots \rangle_N$ indicates an empirical average over N nonstationary Markov chains. We consider the same pool of data as in Sec. IV A. Thus, $N = 6 \times 10^7$, and each trajectory is made up of 500 steps. Accordingly, both the fast-growth and Monte Carlo simulations include 3×10^{10} steps in total, which produces similar running times.

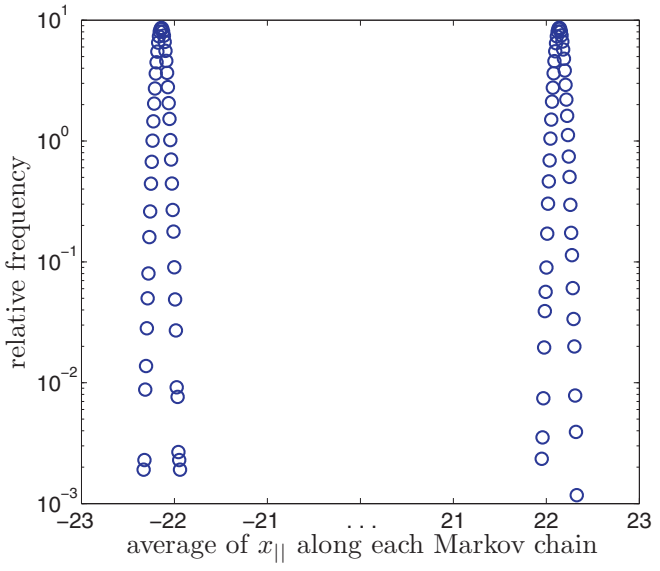

FIG. 7. (Color online) Histogram for the average of $x_\parallel$ along each Markov chain in our standard Monte Carlo program. Despite their different weights, the peaks of the bimodal posterior distribution are sampled almost equally.

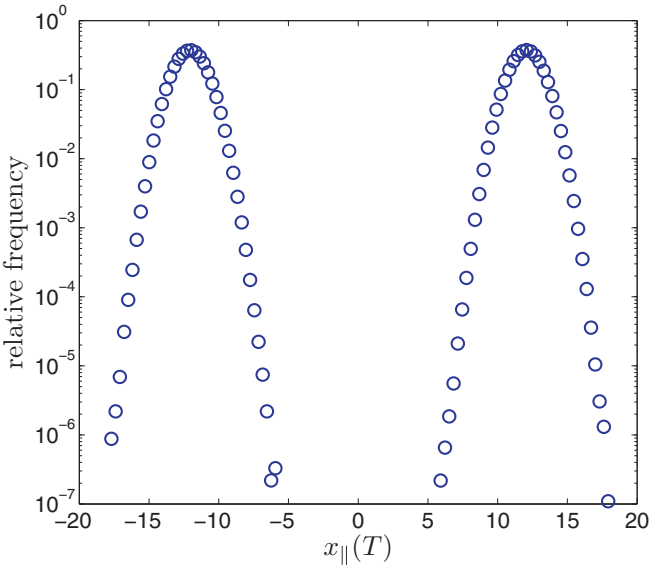

FIG. 8. (Color online) Histogram for the end location $x_\parallel(T)$ of each trajectory in our fast-growth program. Although the Markov chains are nonstationary, they still get trapped around the maxima of the posterior distribution.

The absolute error in the fast-growth calculation,

$$\left|\frac{\langle e^R x_\parallel(T)\rangle_N}{\langle e^R \rangle_N} - \langle x_\parallel \rangle_{\text{post}}\right| \approx 1.19 \times 10^{-3}, \tag{59}$$

demonstrates that the method performs well. As a matter of fact, one obtains the estimate $\langle x_\parallel \rangle_{\text{post}} \approx -20.0296$. Notably, the histogram in Fig. 8 is qualitatively rather similar to the one in Fig. 7. Even if the Markov chains for the fast-growth algorithm are nonstationary, the mismatched peaks of the posterior distribution are sampled equally. However, weighing the final value of $x_\parallel$ with e^R of the respective trajectory in the

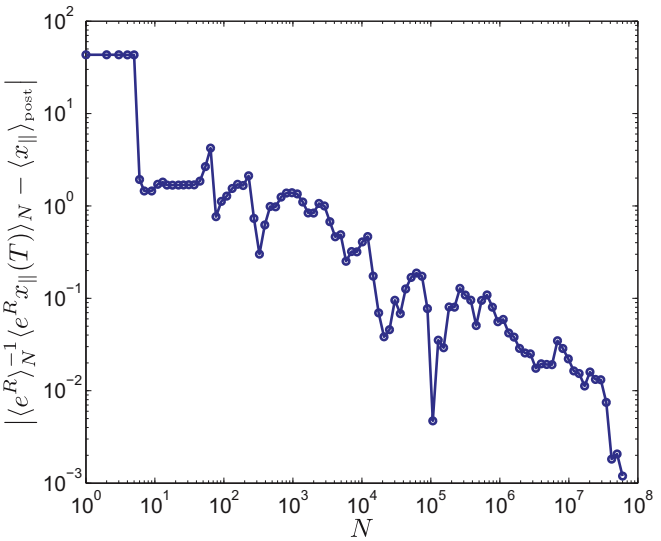

FIG. 9. (Color online) Absolute error in the fast-growth estimate of $\langle x_\parallel \rangle_{\text{post}}$ as the number of trajectories N is varied. The rightmost data point corresponds to (59).

estimator (58) resolves the different weights of the peaks in the posterior distribution and yields the correct result for the average.

Figure 9 specifies the convergence of the fast-growth algorithm as N increases. The detailed error analysis is a topic for future work.

VI. SUMMARY

Successful use of Bayesian methods in realistic problems of statistical data analysis requires efficient ways to numerically calculate high-dimensional integrals. Due to the similarity of this problem with the determination of free-energy differences of complex molecules, the transfer of methods from statistical mechanics to Bayesian statistics has a long tradition. Notably, thermodynamic integration, which replaces the determination of a normalization factor by an integral over much more accessible averages, has proven very valuable in this connection.

However, relying on well-equilibrated averages for different temperatures, thermodynamic integration may run into difficulties in the presence of multimodal distributions. Since multimodal likelihoods and posterior distributions are quite common in Bayesian data analysis, a method less dependent on perfect equilibration is called for. In statistical mechanics, the Jarzynski equation and the Crooks relation have been used successfully to determine free-energy differences from nonequilibrium trajectories without final relaxation. Slightly modified variants of these relations may be implemented to determine the evidence and posterior averages, respectively, in Bayesian statistics.

In the present paper, we have performed a detailed analysis of the statistical error inherent in these methods. From the determination of free-energy differences with the help of the Jarzynski equation, it is known that the method has a bias due to the nonlinearities involved and statistical subtleties of exponential averages. To keep track of these errors in the setting of Bayesian data analysis, we have split the mean-square error of the estimator into a contribution from the bias and from the variance. As usual, the variance may be well characterized by the empirical sample variance, whereas the bias depends on the exact value of the log-evidence $\ln p(d|\mathfrak{M})$, which is not known. We have therefore split the bias once more into a contribution that, similarly to the variance, may be characterized by the sample data alone, and a remainder for which we provide bounds in the form of a confidence interval. Taking everything together, we finally give a confidence interval for the single-block estimate of the log-evidence, $\ln p(d|\mathfrak{M})$, determined from nonstationary Markov chain Monte Carlo simulations.

We have tested our results against extensive numerical simulations of two model cases with bimodal likelihoods. These are either sums of two Gaussians or of two Lorentzians. Combined with appropriate prior distributions, the evidence can be calculated analytically for both cases, which facilitates the comparison with the simulation results. By investigating various samples sizes N, our analytical findings were all verified, and the predicted dependence of the error measures on N was reproduced. Our results are also consistent with error measures discussed previously in connection with free-energy estimates. Similarly, agreement was found for the determination of averages with multimodal posterior distributions using the Crooks relation, where straight Monte Carlo sampling of the posterior was seen to be problematic.

In conclusion, variants of the recently discovered fluctuation theorems of nonequilibrium statistical mechanics may prove very helpful in Bayesian data analysis if multimodal distributions are relevant. In these cases, they allow an efficient determination of high-dimensional integrals via Markov chain Monte Carlo methods without requiring complete equilibration. Admittedly, these methods build upon exponential averages that may converge poorly and that show a bias that needs to be monitored. As in statistical mechanics, the tradeoff between problems of equilibration and subtleties of exponential averages is difficult to assess in general, and it has to be analyzed for each case at hand individually.

ACKNOWLEDGMENTS

Financial support from the Deutsche Forschungsgemeinschaft (DFG) under Project No. EN 278/6-1 is gratefully acknowledged.

[1] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis* (Chapman and Hall, London, 1995).
[2] T. Leonhard and J. S. J. Hsu, *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers* (Cambridge University Press, Cambridge, 1999).
[3] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).
[4] V. Dose, Rep. Prog. Phys. **66**, 1421 (2003).
[5] G. D'Agostini, Rep. Prog. Phys. **66**, 1383 (2003).
[6] *Bayesian Statistics 9*, edited by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Oxford University Press, Oxford, 2011).
[7] U. von Toussaint, Rev. Mod. Phys. **83**, 943 (2011).
[8] P. W. Anderson, Phys. Today **45**(1), 9 (1992).
[9] W. von der Linden, R. Preuss, and V. Dose, in *Maximum Entropy and Bayesian Methods*, edited by W. von der Linden, V. Dose, R. Fischer, and R. Preuss (Kluwer, Dordrecht, 1999).
[10] D. M. Zuckerman and T. B. Woolf, Chem. Phys. Lett. **351**, 445 (2002).
[11] J. Gore, F. Ritort, and C. Bustamante, Proc. Natl. Acad. Sci. (USA) **100**, 12564 (2003).
[12] K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).
[13] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[14] R. M. Neal, Tech. Rep. CRG-TR-93-1, Department of Computer Science, University of Toronto (1993).
[15] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, Oxford, 2006).
[16] H. Ahlers and A. Engel, Eur. Phys. J. B **62**, 357 (2008).
[17] J. K. Stockton, X. Wu, and M. A. Kasevich, Phys. Rev. A **76**, 033613 (2007).
[18] J.-M. Marin and C. Robert, *Bayesian Core: A Practical Approach to Computational Bayesian Statistics* (Springer Science+Business Media, New York, 2007).

[19] N. Chopin, T. Lelièvre, and G. Stoltz, Statist. Comput. **22**, 897 (2011).
[20] C. Jarzynski, Annu. Rev. Condens. Matter Phys. **2**, 329 (2011).
[21] U. Seifert, Rep. Prog. Phys. **75**, 126001 (2012).
[22] M. Esposito, Phys. Rev. E **85**, 041125 (2012).
[23] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, Science **296**, 1832 (2002).
[24] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, Nature (London) **437**, 231 (2005).
[25] A. Pohorille, C. Jarzynski, and C. Chipot, J. Phys. Chem. B **114**, 10235 (2010).
[26] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[27] C. Jarzynski, J. Stat. Mech.: Theory Exp. (2004) P09005.
[28] D. M. Zuckerman and T. B. Woolf, J. Stat. Phys **114**, 1303 (2003).
[29] F. M. Ytreberg and D. M. Zuckerman, J. Comput. Chem. **25**, 1749 (2004).
[30] G. E. Crooks, Phys. Rev. E **61**, 2361 (2000).
[31] D. S. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, Oxford, 1996).
[32] V. Y. Chernyak, M. Chertkov, and C. Jarzynski, J. Stat. Mech.: Theor. Exp. (2006) P08001.
[33] C. Chatelain, J. Stat. Mech.: Theor. Exp. (2007) P04011.
[34] S. R. Williams, D. J. Searles, and D. J. Evans, Phys. Rev. Lett. **100**, 250601 (2008).
[35] P. Billingsley, *Probability and Measure*, Wiley Series in Probability and Mathematical Statistics, 3rd ed. (Wiley-Interscience, New York, 1995).
[36] C. Jarzynski, Phys. Rev. E **56**, 5018 (1997).
[37] D. L. Harnett, *Statistical Methods*, 3rd ed. (Addison-Wesley, Reading, MA, 1982).
[38] R. H. Wood, W. C. F. Mühlbauer, and P. T. Thompson, J. Phys. Chem. **95**, 6670 (1991).
[39] D. S. Sivia and C. J. Carlile, J. Chem. Phys. **96**, 170 (1992).
[40] T. Appourchaux, L. Gizon, and M.-C. Rabello-Soares, Astron. Astrophys., Suppl. Ser. **132**, 107 (1998).
[41] M. Gruberbauer, T. Kallinger, W. W. Weiss, and D. B. Guenther, Astron. Astrophys. **506**, 1043 (2009).