Bayesian generalised ensemble Markov chain Monte Carlo

Jes Frellsen
University of Cambridge

Ole Winther
Technical University of Denmark

Zoubin Ghahramani
University of Cambridge

Jesper Ferkinghoff-Borg
University of Copenhagen

Abstract

Bayesian generalised ensemble (BayesGE) is a new method that addresses two major drawbacks of standard Markov chain Monte Carlo algorithms for inference in high-dimensional probability models: inapplicability to estimate the partition function and poor mixing properties. BayesGE uses a Bayesian approach to iteratively update the belief about the density of states (distribution of the log likelihood under the prior) for the model, with the dual purpose of enhancing the sampling efficiency and making the estimation of the partition function tractable. We benchmark BayesGE on Ising and Potts systems and show that it compares favourably to existing state-of-the-art methods.

1 INTRODUCTION

For most probabilistic models, $p(\mathbf{x})$, exact inference is intractable, and one has to resort to some form of approximation (Bishop 2006). Markov chain Monte Carlo (MCMC) methods constitute a particularly powerful and versatile approach for this purpose (Metropolis and Ulam 1949; Metropolis et al. 1953; Hastings 1970). In standard MCMC methods, $p(\mathbf{x})$ is sampled through a Markov chain with a fixed transition kernel that is constructed to have the unique invariant distribution $p(\mathbf{x})$. In the following we will broadly refer to this as *canonical* sampling.

While the canonical MCMC method has been the workhorse in statistics and physics for the last 50 years or so, it suffers from two primary deficiencies. First, it only samples from a narrow interval of log-likelihood values, which makes the method inapplicable for calculating key multivariate integrals, in particular the *partition function* (*evidence*, *marginal likelihood*) or the *density of states* associated with $p(\mathbf{x})$ (Iba 2001; Bishop 2006; Ferkinghoff-Borg 2012). Secondly, canonical sampling is often hampered by a high degree of correlations between the generated states for standard choices of the transition kernel. This property, which is referred to as *poor mixing* or *slow relaxation* of the Markov chain, reduces the effective number of samples and may lead to results which are erroneously sensitive to the arbitrary initialisation of the chain.

In the past few decades, a variety of MCMC methods known as extended ensembles have been proposed to alleviate these two deficiencies (see review and reference in Gelman and Meng (1998), Iba (2001), Murray (2007), and Ferkinghoff-Borg (2012)). The underlying idea of this approach is to build a "bridge" from the part of the probability distribution where the Markov chain suffers from slow relaxation to the part where the sampling is free from such problems. These methods include simulated tempering (Marinari and Parisi 1992; Lyubartsev et al. 1992; Irbäck and Potthast 1995), parallel tempering (Swendsen and Wang 1986; Geyer 1991) and *generalised ensembles* (Berg and Neuhaus 1992; Lee 1993; Hesselbo and Stinchcombe 1995). In extended ensembles the transition kernel is extended in such way that it at the same time allows for the calculation of the partition function as well as for the reconstruction of the desired statistics for the original target distribution $p(\mathbf{x})$. However this kernel depends on integral quantities of the model which are not *a priori* known. Therefore these techniques rely on an iterative approach, where estimates of these quantities obtained from previous iteration(s) are used to define the transition probability kernel for the next iteration (Murray 2007; Ferkinghoff-Borg 2012).

At present, however, extended ensemble techniques use rather heuristic approaches to define the required iteration procedure and are all based on frequentist estimators. Consequently, whilst they have shown

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

promising results in a wide range of problems in machine learning and statistical physics (Salakhutdinov 2010; Landau and Binder 2014), these aspects compromise both the robustness and speed of the algorithms and limit their applicability for inference in more complex systems.

In this paper we focus on the generalised ensemble (GE) subclass of methods, due to its larger domain of application compared to the tempering based counterparts (Hansmann and Okamoto 1997). We propose a novel Bayesian generalised ensemble (BayesGE) method to address the problems pertaining the traditional GE techniques. The proposed method is applicable to both discrete and continuous distributions but requires discretisation of the log-likelihood in the current formulation. We test the algorithm on two discrete 2D spin systems: an Ising model and a Potts model. Both models are canonical examples of systems displaying cooperative transitions and slow relaxation, for which tempering based inference methods in general are inadequate. Inference in Potts models is further complicated by their generic multimodal nature as opposed to the essential bimodal nature of Ising systems. We demonstrate the robustness and accuracy of the algorithm in estimating the full density of states and partition function of the two models for different size and/or complexity and show that BayesGE outperforms existing state-of-the-art methods: the *Wang–Landau* (WL) algorithm (Wang and Landau 2001), *annealed importance sampling* (AIS) (Neal 2001) and *nested sampling* (Skilling 2006; Murray et al. 2005).

In section 2, we outline the basic methodology of generalised ensembles. In section 3, we detail the elements of the BayesGE method. The inference results of the algorithm in comparison to existing advanced sampling methods on the two spin systems are presented in section 4. We conclude by discussing relevant extensions of the method for future work.

2 GENERALISED ENSEMBLES

Consider a posterior distribution of $\mathbf{x} \in \mathcal{X}$ on the form

$$p_\beta(\mathbf{x}) = \frac{\exp(-\beta\mathcal{E}(\mathbf{x}))p_0(\mathbf{x})}{Z_\beta} \tag{1}$$

where

$$Z_\beta = \int \exp(-\beta\mathcal{E}(\mathbf{x}))p_0(\mathbf{x})\,\mathrm{d}\mathbf{x}\ . \tag{2}$$

In the context of statistical physics, β represents the inverse temperature times the Boltzmann constant, $\mathcal{E}(\mathbf{x})$ is the energy and the normaliser Z_β is known as the *partition function*, which plays a central role due to its link to the thermodynamical free energy. Here, $p_0 = p_{\beta=0}$ represents the normalised integration measure and consequently $Z_0 \equiv Z_{\beta=0} = 1$.[1] In the non-thermal context of Bayesian statistics, we set $\beta = 1$ in which case the "energy" equals minus log likelihood $\mathcal{E}(\mathbf{x}) = -\log p(\text{data}|\mathbf{x})$, $\mathbf{x}$ are model parameters and latent variables, p_0 is the prior distribution and $Z \equiv Z_{\beta=1}$ is known as the *model evidence* or *marginal likelihood*. Another central object is the *density of states* which is defined as

$$g(E) = \int \delta(E - \mathcal{E}(\mathbf{x}))p_0(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3}$$

and measures the prior mass associated with energy E. In statistical physics, $g(E)$ is related to the *microcanonical entropy* $s(E)$ through Boltzmann's formula $s(E) \equiv \log g(E)$, from which all thermodynamic potentials can be calculated. Typically, the energy can be evaluated for any $\mathbf{x}$ but the partition function Z_β and g are unknown.

A Markov chain with unique invariant distribution p_β can be constructed using the Metropolis–Hastings (MH) algorithm (Metropolis et al. 1953; Hastings 1970), in which a sequence of states $\{\mathbf{x}_i\}$ is generated by sampling a trial state from a proposal distribution $\mathbf{x}' \sim q(\cdot|\mathbf{x}_i)$ and accepting the state with probability $a(\mathbf{x}'|\mathbf{x}_i) = \min\left\{1, \frac{p_\beta(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p_\beta(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\right\}$ at each time step. Typically, p_β only occupies an exponentially small part of the volume under p_0, as illustrated in figure S1. For the Metropolis–Hastings algorithm, this implies that the normalisation constant Z_β (as well as other multivariate-integrals) is not tractable and furthermore that the sampling for more complex models may suffer from slow mixing of the Markov chain.

In the GE procedure, an artificial target distribution is constructed that facilitates a "bridging" between p_0 and p_β. The starting point is to replace the log-Boltzmann weights $-\beta E$ with a different weight function $w(\mathbf{y}(\mathbf{x}))$, where $\mathbf{y}(\mathbf{x})$ represent a set of functions of $\mathbf{x}$ of particular interest ("reaction coordinates" or "order parameters"). The most common choice is to set $y = \mathcal{E}$ in which case the GE target distribution takes the form

$$p_{\mathrm{GE}}(\mathbf{x}) = p(\mathbf{x}|w) \equiv \frac{\exp(w(\mathcal{E}(\mathbf{x})))p_0(\mathbf{x})}{Z_w} \tag{4}$$

where

$$Z_w = \int \exp(w(\mathcal{E}(\mathbf{x})))p_0(\mathbf{x})\,\mathrm{d}\mathbf{x}\ . \tag{5}$$

[1] Note, that the physical partition function Z_0 is normally identified with the total volume of $\mathcal{X}$, corresponding to p_0 being an actual integration measure as opposed to a normalised one. However, for most statistical considerations it is the ratio Z_β/Z_0 that is of primary importance, so this normalisation convention poses no loss of generality.

Sampling from $p(\mathbf{x}|w)$ is realised with MH by replacing $p_\beta(\mathbf{x})$ with $p(\mathbf{x}|w)$ in $a(\mathbf{x}'|\mathbf{x}_i)$.

In order to ensure the target distribution has the desired properties, GE methods make use of the density of states to define the weights. Using the definition from equation (3), the marginal distribution of the energy $p(E|w) = \int \delta(E - \mathcal{E}(\mathbf{x}))p(\mathbf{x}|w)\,\mathrm{d}\mathbf{x}$ can be expressed as

$$p(E|w) = \frac{\exp(w(E) + s(E))}{Z_w} \tag{6}$$

and the partition function from equation (5) as

$$Z_w = \int \exp(w(E) + s(E))\,\mathrm{d}E\ . \tag{7}$$

We can use this distribution to estimate canonical quantities for any choice of β, for example

$$\begin{aligned} Z_\beta &= Z_w \int \exp(-\beta E - w(E))p(E|w)\,\mathrm{d}E \\ &= \int \exp(-\beta E + s(E))\,\mathrm{d}E\ . \end{aligned} \tag{8}$$

So in principle, for any choice of $w(E)$ we can collect samples from $p(\mathbf{x}|w)$ and thereby from $p(E|w)$ to get posterior estimates.

If we know the entropy $s(E)$ we can then define two prominent generalised ensembles: the *multicanonical ensemble* (MUCA) (Berg and Neuhaus 1992; Lee 1993) and the $1/k$ *ensemble* (Hesselbo and Stinchcombe 1995) with weights given respectively by

$$w_{\mathrm{MUCA}}(E) = -s(E) \qquad w_{1/k}(E) = -\log k(E) \tag{9}$$

where $k(E) \equiv \int_{-\infty}^{E} g(E')\,\mathrm{d}E'$. The multicanonical ensemble makes the marginal distribution $p(E|w)$ flat, which implies that once the Markov chain has converged all energies are visited with equal probability. The $1/k$ ensemble is constructed to put roughly equal probability to all values of the entropy, which leads to more frequent sampling of the low-energy part compared to the multicanonical sampling, c.f. section S1.

The primary obstacle of the GE approach is that $s(E)$ is not known *a priori*. Indeed, had this been the case the partition function Z_β could have been directly calculated according to the one-dimensional integral (8). The fact that the GE ensemble is defined in terms of quantities that are the primary aim to infer from the method in the first place, implies that all GE methods rely on a iterative procedure in which weights are iteratively refined based on the data and weights from previous iteration(s). The different GE learning algorithms differ in the choice of data, estimators and iteration procedure, but are all based on count statistics and do not account for prior knowledge in any systematic manner, c.f. section S2.

3 BAYESIAN GENERALISED ENSEMBLE

We wish to address the problems of traditional GE approaches in a principled manner, by treating the entropy estimation as an inference problem in a Bayesian framework. In the following we will assume that we are considering a discrete or discretised system having J possible energy values. To this end, let $\mathbf{s} = (s_1, \ldots, s_J)$ be the entropy vector over the energy values $\mathbf{E} = (E_1, \ldots, E_J)$ with $s_j = \log g(E_j)$.

Assume that for each weight vector $\mathbf{w}^{(\tau)} = (w_1, \ldots, w_J)$ in a given set of weights $W = \{\mathbf{w}^{(\tau)}\}_{\tau=1}^{t}$ we perform a Metropolis–Hastings simulation with $\nu^{(\tau)}$ steps and target distribution

$$P\big(\mathbf{x}|\mathbf{w}^{(\tau)}\big) = \frac{\exp w_j^{(\tau)}}{Z_{\mathbf{w}^{(\tau)}}} P_0(\mathbf{x})\ , \tag{10}$$

where $\mathcal{E}(\mathbf{x}) = E_j$. Based on the samples $\{\mathbf{x}_i^{(\tau)}\}_{i=1}^{\nu^{(\tau)}}$ from the τ'th simulation, we can compute an energy histogram $\mathbf{n}^{(\tau)} = (n_1^{(\tau)}, \ldots, n_J^{(\tau)})$, where $n_j^{(\tau)} = |\{\mathbf{x}_i^{(\tau)}|\mathcal{E}(\mathbf{x}_i^{(\tau)}) = E_j\}|$. This gives us a set of t energy histograms $N = \{\mathbf{n}^{(\tau)}\}_{\tau=1}^{t}$. Based on these t simulations we can write the posterior distribution of the entropy $\mathbf{s}$ as

$$P(\mathbf{s}|N, W, \sigma) = \frac{P(N|W, \mathbf{s})P(\mathbf{s}|\sigma)}{P(N|W, \sigma)}\ , \tag{11}$$

where σ are hyperparameters of the prior. We propose to use a Gaussian process prior for $\mathbf{s}$ and a product of multinomial distributions for the likelihood of N. As detailed below and in algorithm 1, this allows us to define an efficient and robust algorithm, where the posterior $P(\mathbf{s}|N, W, \sigma)$ is iteratively updated and used to define the weights $\mathbf{w}^{(t+1)}$ for the next simulation.

3.1 Prior Specification

For continuous systems $s(E)$ is typically a smooth and (mostly) concave function with a non-trivial shape. Similarly, for a discrete or discretised system $\mathbf{s}$ will be values of such a function. We propose to use a Gaussian process prior $\mathcal{GP}(0, \kappa)$ for $s(E)$ with a suitable kernel κ (Rasmussen and Williams 2006). As $s(E)$ is typically concave, the kernel needs to be non-stationary because the entropy should not revert back to the mean function away from observed energies. Concavity cannot directly be modelled on quadratic form, but a cubic spline extrapolates linearly and gives a quite flexible fit. Linear extrapolation appears to be a good choice because we need be reasonably conservative in order to keep the iterative estimation of weights robust.

The cubic spline prior penalises curvature, and by assuming the boundary conditions $s(0) = s'(0) = 0$ the kernel on the unit interval can be expressed as (section 6.3.1 in Rasmussen and Williams 2006; Wahba 1978)

$$\kappa(E, E'|\sigma) = \sigma^2 \left(|E - E'|v^2/2 + v^3/3\right) , \qquad (12)$$

where $v = \min(E, E')$. As these boundary conditions are not appropriate for our problem, we remove them by incorporating explicit linear basis functions $\mathbf{u}(E) = (1, E)^\intercal$ with a normal prior on the coefficients $\mathcal{N}(\mathbf{0}, B)$, $B \in \mathbb{R}^{2\times 2}$. By integrating out the coefficients we obtain the prior

$$s|\sigma \sim \mathcal{GP}(0, \kappa(E, E'|\sigma) + \mathbf{u}(E)^\intercal B \mathbf{u}(E')) , \qquad (13)$$

where we take the limit $B^{-1} \to 0$ to make the prior on the basis functions vague (section 2.7 in Rasmussen and Williams 2006).

3.2 Likelihood

The probability of observing a histogram $\mathbf{n} \in N$ generated with the weights $\mathbf{w} \in W$ is naturally expressed through the multinomial distribution (Ferkinghoff-Borg 2002)

$$p(\mathbf{n}|\mathbf{w}, \mathbf{s}) = \nu! \prod_{j\in\mathcal{S}} \frac{p(E_j|\mathbf{w}, \mathbf{s})^{n_j}}{n_j!} , \qquad (14)$$

where $\nu = \sum_{j\in\mathcal{S}} n_j$, $\mathcal{S}$ is the index set associated with the histogram (see below) and

$$p(E_j|\mathbf{w}, \mathbf{s}) = \frac{\exp(w_j + s_j)}{Z_\mathbf{w}} , \qquad (15)$$

where

$$Z_\mathbf{w} = \sum_{j\in\mathcal{S}} \exp(w_j + s_j) . \qquad (16)$$

Equation (15) is the discrete version of equation (6). If $\mathbf{n}$ represents a fully equilibrated sample from $p(\mathbf{x}|\mathbf{w})$ the sum in equation (16) should run over the full index set $\mathcal{S} = \{1, \ldots, J\}$. However, if we do not have a fully equilibrated sample, it was proposed by Ferkinghoff-Borg (2002) to assume that each histogram $\mathbf{n}$ only represents an equilibrated sampling within the observed support $\tilde{\mathcal{S}} = \{j \mid n_j > 0\}$. In this case we would constrain the summation in the normalisation constant $Z_\mathbf{w}$ to $\tilde{\mathcal{S}}$ to reflect the assumption of local equilibration only, as detailed by Ferkinghoff-Borg (2002, 2012). As the weights and entropy function render the histograms independent, the probability for the combined set of observations N is then given by

$$P(N|W, \mathbf{s}) = \prod_{\tau=1}^{t} P(\mathbf{n}^{(\tau)}|\mathbf{w}^{(\tau)}, \mathbf{s}) . \qquad (17)$$

To compensate for the likelihood not scaling correctly with the number of *independent* samples, we scale the histograms $\mathbf{n}^{(\tau)}$ with the inverse of the *number of degrees of freedom* of the sampled model.

3.3 Posterior Inference

In order to make posterior inference analytically tractable, we approximate the log-likelihood function from equation (17) by a second order Taylor expansion. Using the maximal likelihood estimate (MLE) $\hat{\mathbf{s}} = \arg\max_\mathbf{s} P(N|W, \mathbf{s})$ as expansion point this approximation reads

$$\log P(N|W, \mathbf{s}) \approx \log P(N|W, \hat{\mathbf{s}}) - \frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^\intercal H (\mathbf{s} - \hat{\mathbf{s}}) , \qquad (18)$$

where $\hat{s}_j$ is only well-defined for $j \in S = \cup_{\tau=1}^{t} \tilde{\mathcal{S}}^{(\tau)}$. H is given by the negative Hessian

$$\begin{aligned} H &= -\frac{\partial^2 \log P(N|W, \mathbf{s})}{\partial \mathbf{s}^2}(\hat{\mathbf{s}}) \\ &= \sum_{\tau=1}^{t} \nu^{(\tau)} \left[\mathrm{diag}(\hat{\mathbf{p}}^{(\tau)}) - \hat{\mathbf{p}}^{(\tau)} \cdot (\hat{\mathbf{p}}^{(\tau)})^\intercal\right] , \end{aligned} \qquad (19)$$

where $\hat{p}_j^{(\tau)} = P(E_j|\mathbf{w}^{(\tau)}, \hat{\mathbf{s}})$. We note that third (and higher) order terms in the expansion are generally negligible as these involve two (or more) outer products of $\hat{\mathbf{p}}^{(\tau)}$ and for the broad target distribution of GE methods $\left\|\hat{\mathbf{p}}^{(\tau)}\right\|_\infty \ll 1$. The MLE $\hat{\mathbf{s}}$ solution can be found using the *generalised multi-histogram* (GMH) equations (Ferkinghoff-Borg 2002), which involve a set of t nonlinear equations (S3) with unknowns $\{Z_{\mathbf{w}^{(\tau)}}\}_{\tau=1}^{t}$. These can be solved effectively using the standard iterative Newton–Raphson method, and $\hat{\mathbf{s}}$ can be calculated from the solution using equation (S4), see section S3 for details.

Using this second order approximation, $P(N|W, \mathbf{s}) \approx \mathcal{N}(\hat{\mathbf{s}}, H)$, we can perform posterior inference analytically, as both the prior and likelihood are Gaussian. This means that $P(\mathbf{s}|N, W, \sigma) \approx \mathcal{N}(\bar{\mathbf{s}}, V)$ and when taking the limit $B^{-1} \to 0$ in equation (13) we get (section 2.7 in Rasmussen and Williams 2006)

$$\bar{\mathbf{s}} = \kappa(\mathbf{E}, \mathbf{E}_S|\sigma) K_H^{-1} \hat{\mathbf{s}} + R^\intercal (U K_H^{-1} U^\intercal)^{-1} U K_H^{-1} \hat{\mathbf{s}} \qquad (20)$$

$$\begin{aligned} V = {} & \kappa(\mathbf{E}, \mathbf{E}|\sigma) - \kappa(\mathbf{E}, \mathbf{E}_S|\sigma) K_H^{-1} k(\mathbf{E}_S, \mathbf{E}) \\ & + R^\intercal (U K_H^{-1} U^\intercal)^{-1} R \end{aligned} \qquad (21)$$

where $\mathbf{E}_S = (E_j | j \in S)$, $H_S = [H_{jj'} | j, j' \in S]$, $K_H = \kappa(\mathbf{E}_S, \mathbf{E}_S|\sigma) + H_S^{-1}$, $U = \mathbf{u}(\mathbf{E})$, $U_S = \mathbf{u}(\mathbf{E}_S)$ and $R = U_S - U K_H^{-1} \kappa(\mathbf{E}_S, \mathbf{E}|\sigma)$. An estimate $\hat{\sigma}$ for the hyperparameter is obtained by optimising the marginal likelihood $P(N|W, \sigma)$ w.r.t. σ. See section S4 for further details.

We have also tested other approximations to the posterior, including a quadratic approximation to the individual likelihood terms from equation (14) and a Laplace approximation to the posterior. Approximating the individual terms turned out to be too inaccurate. For the Laplace approximation we can update

the posterior either sequentially using the current histogram plus the previous Laplace approximation, or using all histograms plus the Gaussian process prior. The latter will in principle give the best approximation, however the Laplace approximation involves solving a $|S|$ dimensional optimisation problem, while the MLE can be found by solving a system of t nonlinear with t unknown, where $t \ll |S|$ typically.

3.4 Setting the Weights

At each iteration of the algorithm, we can use the posterior mean estimate $\bar{\mathbf{s}}$, equation (20), to define the weights of the generalised ensemble in the next step. The simplest approach is to use the definition of the GE ensemble from equation (9) directly which implies the updating rules for respectively multicanonical and $1/k$ weights:

$$\bar{\mathbf{w}}[\text{MUCA}]^{(t+1)} = -\bar{\mathbf{s}} \tag{22}$$

$$\bar{w}[1/k]_j^{(t+1)} = -\log \sum_{j' \leq j} \exp(\bar{s}_{j'}) \ . \tag{23}$$

However, this rule fails to account for the uncertainty of the posterior of $\mathbf{s}$. A more robust approach is to define the weights according to *expectation values* of the marginal distribution $P(E_j|\mathbf{w}^{(t+1)})$ under the posterior distribution $P(\mathbf{s}|N, W, \hat{\sigma})$, where N and W are the set of histograms and weights for the first t iterations. For any weight function $\mathbf{w}$ this expectation is given as

$$\langle P(E_j|\mathbf{w})\rangle_{P(\mathbf{s}|N,W)} = \left\langle \frac{\exp(w_j + s_j)}{Z_{\mathbf{w}}} \right\rangle_{P(\mathbf{s}|N,W)} \tag{24}$$

In section S5 we show that, to first order in $\delta\mathbf{s} = \mathbf{s} - \bar{\mathbf{s}}$, the expected GE target distribution is realised by simply setting

$$\tilde{\mathbf{w}}^{(t+1)} = \bar{\mathbf{w}}^{(t+1)} - \frac{1}{2}\operatorname{diag}(V) \ , \tag{25}$$

where $\bar{\mathbf{w}}^{(t+1)}$ are the weights defined in equations (22) to (23) and V is the posterior covariance matrix (21).

3.5 Algorithmic Details

The BayesGE algorithm is outlined in algorithm 1. We use a simple exponential scheme for setting the simulation time for each iteration of the algorithm, which has previous been shown to be efficient (Ferkinghoff-Borg 2002). The simulations time $\nu^{(t+1)}$ for the $(t+1)$'th iteration of the algorithm is set to

$$\nu^{(t+1)} = \begin{cases} \gamma\nu^{(t)} & \text{if } \mathbf{n}^{(t)} \text{ and } \sum_{\tau=1}^{t-1} \mathbf{n}^{(\tau)} \\ & \text{have the same support} \\ \nu^{(t)} & \text{otherwise} \end{cases} \ , \tag{26}$$

Algorithm 1 The BayesGE algorithm

Input: $\nu^{(1)}$, γ

1: $\mathbf{w}^{(1)} \leftarrow \mathbf{0}$
2: $N, W \leftarrow \{\}, \{\}$
3: **for** $t \in (1, \ldots, T)$ **do**
4: Draw $\nu^{(t)}$ samples $\{\mathbf{x}_i\}_{i=1}^{\nu^{(t)}} \overset{\text{M.H.}}{\sim} P(\mathbf{x}|\mathbf{w}^{(t)})$
5: Compute an energy histogram $\mathbf{n}^{(t)}$ based on the samples $\{\mathbf{x}_i\}_{i=1}^{\nu^{(t)}}$
6: Insert $\mathbf{n}^{(t)}$ and $\mathbf{w}^{(t)}$ into N and W respectively
7: Optimise $P(N|W, \sigma)$ w.r.t. σ and update the posterior $P(\mathbf{s}|N, W, \hat{\sigma})$
8: Set the weight for next iteration $\mathbf{w}^{(t+1)}$ using $P(\mathbf{s}|N, W, \hat{\sigma})$ and equation (25)
9: Set the simulation time for the next iteration $\nu^{(t+1)}$ using equation (26)

Output: $P(\mathbf{s}|N, W, \hat{\sigma})$

where the increasement factor $\gamma > 1$. The motivation for this scheme is that if we did not see any new energies in the last simulations then we either need to run longer simulations to see new energies or we have seen all possible energies. If we have seen new energies in the last simulations, then the simulation time is already long enough to explore new energies. As default values we use $\nu^{(1)} = 5000$ and $\gamma = 2^{1/10}$. The algorithm is robust to changes in these values and they do not change the performance of the algorithm significantly.

3.6 Complexity and Convergence

The computational complexity of the BayesGE algorithm stems from $\nu = \sum_{\tau=1}^{t} \nu^{(\tau)}$ evaluations of the posterior target distribution, equation (1), and $T = \mathcal{O}(\log \nu)$ entropy inference steps each of which is $\mathcal{O}(J^3)$. Accordingly, in term of MC steps ν the algorithm scales as $\mathcal{O}(\nu c + J^3 \log \nu)$, where c is the complexity of evaluating the posterior target, equation (1). The latter term can be reduced to be linear in J using sparse GP approximations (Quiñonero-Candela and Rasmussen 2005). Note that the BayesGE algorithm does not introduce an additional cost at the individual MC step, as the weights are kept fixed for each iteration of the algorithm. For fully connected models (e.g. Boltzmann machines and polymer systems) the complexity of posterior target evaluations scales with the square of the number of degrees of freedom and linear with the number of data points, in which case the overhead of the inference step is expected to be negligible, as empirically illustrated in section S7.

Convergence to the target ensemble is in principle guaranteed by the formal correctness of the inference procedure (in keeping with the preservation of de-

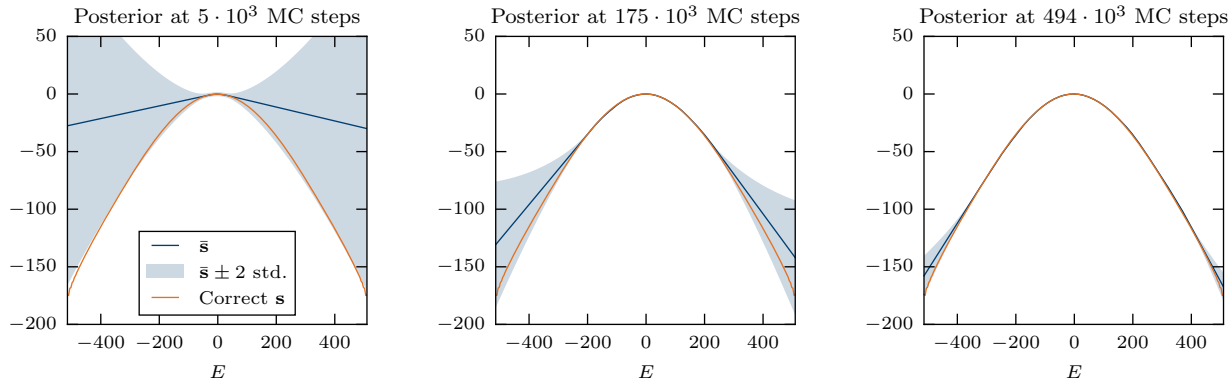

Figure 1: Examples of the posterior distribution of the entropy function $\mathbf{s}$ for BayesGE with multicanonical weight at three different simulations lengths (corresponding to 1, 25 and 45 histograms) for the 2D Ising model of size 16×16. The orange line is the ground truth and the blue line is the posterior mean estimate $\bar{\mathbf{s}}$, equation (20), with the shaded area showing $\pm$ two standard deviations.

tailed balance) and the exponential update rule, equation (26), which serves the dual purpose of ensuring accumulation of statistics as well as asymptotic unbiased sampling. As demonstrated on the chosen models the BayesGE algorithm has a learning stage where the error of an estimator scales with sub $\mathcal{O}(1/\sqrt{\nu})$ and a sampling stage where the error displays a $\mathcal{O}(1/\sqrt{\nu})$ scaling, similar to the WL algorithm (Iba 2001).

4 RESULTS

We have applied the method on two discrete Markov random field model: the 2D square lattice Ising model and Potts model (see Wu (1982) for a review) and compared it against three other state-of-the-art algorithms: the WL algorithm, AIS and nested sampling.

In the Ising model the state is $\mathbf{x} \in \{-1, 1\}^{L^2}$ and in the Potts model $\mathbf{x} \in \{1, 2, \ldots, q\}^{L^2}$, where L is the size of the lattice side and q is the number of colours. We use a homogeneous coupling constant of 1 and no external field in which case the energy of a state is given respectively by

$$\mathcal{E}_{\text{Ising}}(\mathbf{x}) = -\sum_{\langle a,b \rangle} x_a x_b \tag{27}$$

$$\mathcal{E}_{\text{Potts}}(\mathbf{x}) = -\sum_{\langle a,b \rangle} \delta(x_a, x_b) \ , \tag{28}$$

where the sums run over neighbour pairs $\langle a, b \rangle$ in the lattice and δ is the Kronecker delta. For these models we want to estimate the partition function Z_β from equation (2) with a uniform prior.[2] The Ising model has a second order phase transition, while the Potts model has a first order phase transition and is notoriously hard to sample for temperature based methods (Iba 2001). For all algorithms we use a single spin flip Metropolis proposal.

4.1 Posterior Distribution of Entropy

Examples of the posterior distribution of the entropy $\mathbf{s}$ for a simulation of the 16×16 Ising model using BayesGE with multicanonical weights are illustrated in figure 1 and compared to the correct values obtained by analytical means (Beale 1996). We see that the BayesGE algorithm initially learns about the entropy around the peak of the curve (corresponding to low β) and is very uncertain about the tails. As more samples are collected the uncertainty drops in the tails of the entropy function. Similar plots for BayesGE with $1/k$ weights are shown in figure S2. A notable difference is that the multicanonical weights explores the whole phase space, whereas $1/k$ explores the left branch of the entropy curve only. As the Z_β values that we will consider below all correspond to sums with the majority weight on the left branch this give $1/k$ a natural advantage over the multicanonical weights. AIS and nested sampling also essentially only sample the left branch of the entropy curve.

4.2 Estimation of Partition Functions

In figure 2 we compare the performance of BayesGE to WL, AIS and nested sampling (see alternative illustrations in figures S3 and S4) in estimating the partition functions for Ising and Potts models. For each model we ran 50 independent simulations using each method,

[2] The total number of states in the two models are easily calculated by $Z_0 = 2^{L^2}$ and $Z_0 = q^{L^2}$ respectively, using the natural (unnormalised) counting measure for p_0. This implies that we in these cases could obtain *absolute* estimates for any partition function Z_β in the BayesGE methodology.

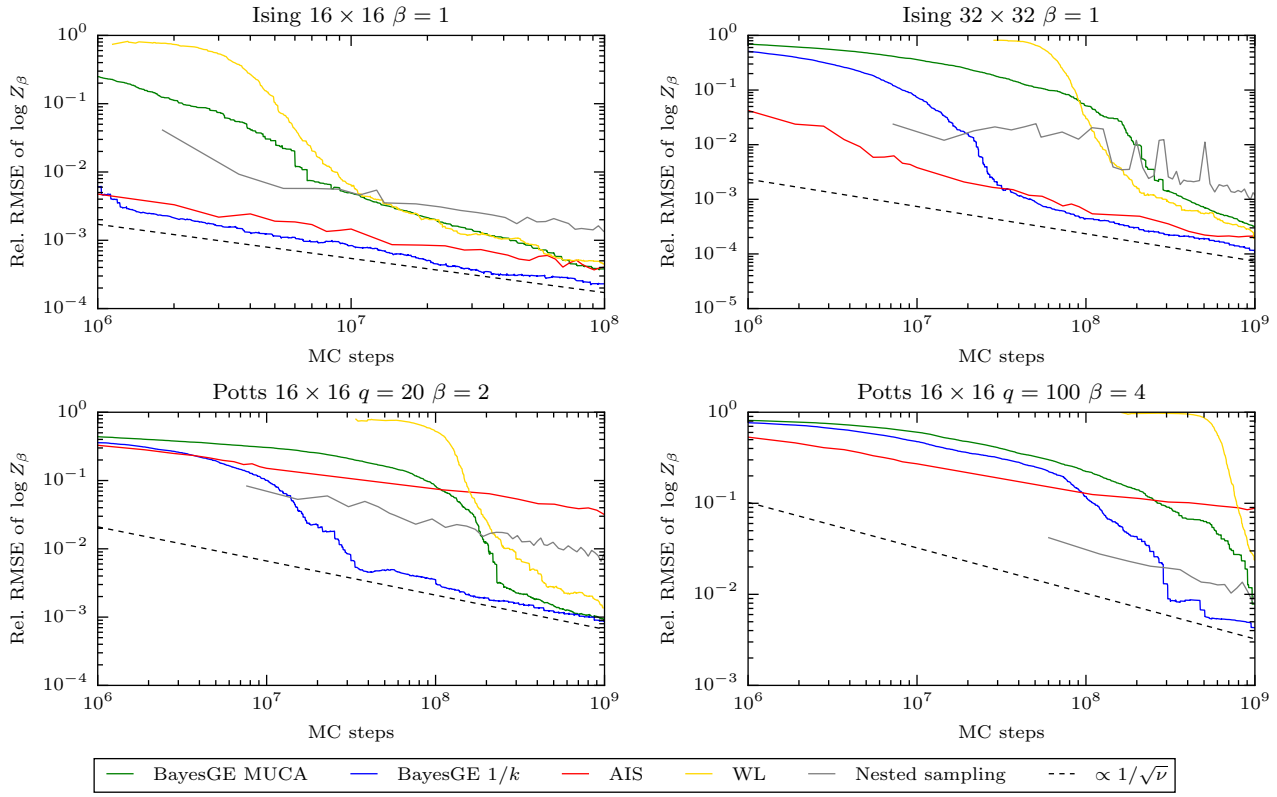

Figure 2: The relative root mean square error (RMSE) of $\log Z_\beta$ as a function of the number Monte Carlo (MC) steps for simulations on 2D Ising and Potts models with different sizes, number of colours (q) and values of β. For an unbiased MC estimator the RMSE scales with the number of MC steps ν as $\mathcal{O}(1/\sqrt{\nu})$, which is indicated with a dashed line. We show results for BayesGE with the multicanonical ensemble (BayesGE MUCA) and the $1/k$ ensemble (BayesGE $1/k$), annealed importance sampling (AIS), the Wang and Landau (WL) algorithm and nested sampling.

and the plots show the relative root mean square error[3] (RMSE) of $\log Z_\beta$ as a function of the number of Monte Carlo (MC) steps. For the Ising model the reference $\log Z_\beta$ is calculated from the analytically entropy function (Beale 1996). For the Potts model we cannot compute $\log Z_\beta$ analytically, so as reference we use the average $\log Z_\beta$ over 50 independent WL simulations using 10^{10} MC steps. For each model, we selected β such that the mode of $p(E|\mathbf{w}_\beta)$ lies well below the phase transition.

We ran BayesGE with multicanonical and $1/k$ weights using the standard settings from section 3, and calculated Z_β using the posterior mean estimate $\bar{\mathbf{s}}$ and equation (8). The implementation details of BayesGE and the other methods are discussed in section S6. The unknown hyperparameters of nested sampling have been optimised prior to the simulation results presented in figure 2 in a laborious trial-and-error manner (see figure S5). Furthermore, both AIS and nested sampling require full rerun for each choice of the total allocated simulation time. In contrast, the GE algorithms simply proceed with the iteration process, and as such provide a more flexible framework for increasing the precision of the estimators.

The first observation to make from figure 2 is that the BayesGE algorithm in all cases display a sigmoidal type of behaviour with a sharp decrease at intermediate times (learning stage) followed by a $\mathcal{O}(1/\sqrt{\nu})$ behaviour at large times (sampling stage), as previously discussed. This long time scaling behaviour testifies to the convergence properties of the algorithm.

If we compare the two multicanonical methods, BayesGE MUCA and WL, we see that they are somewhat on par on the two Ising models. On the Potts models however, BayesGE MUCA does perform much better than WL, with a particular pronounced speedup on the 100 colour problem. This testifies to the advan-

[3]The relative RMSE is $\sqrt{\langle(\log \hat{Z} - \log Z_{\text{ref}})^2\rangle}/\log Z_{\text{ref}}$ where $\hat{Z}$ is the estimated partition function, Z_{ref} the reference partition function and the expectation $\langle\cdot\rangle$ is taken over repeated simulations.

tage of the Bayesian approach, since the target ensemble is the same for the two algorithms.

On all four models AIS performs well at a low number of MC steps, and for the Ising models AIS is one of the best methods. However, for the Potts models, which have a first order phase transition, we see that AIS converges very slowly and performs the worst of all methods. This fits with the expectation that temperature based methods have notorious difficulties on systems with first order phase transitions (Iba 2001), due to trapping in metastable states. Nested sampling also performs well at low number of MC steps, but on the Ising models it has the highest RMSE at the end of the simulations. As expected, nested sampling does not have the same problems as AIS with the first order phase transition in the Potts model. On the 20 colour Potts model the error is about an order of magnitude lower for nested sampling than for AIS at the same number of iterations. However, at the end of the simulation WL and the two BayesGE methods have a lower error than nested sampling. Nested sampling performs very well on the 100 colour Potts model and it is only the two BayesGE methods that has a lower RMSE at 10^9 MC steps.

BayesGE $1/k$ performs consistently well on all four models and in all cases it has the lowest RMSE at the maximal number of MC steps. On the Ising models BayesGE $1/k$ has a similar performance as AIS, though AIS has a lower RMSE at low numbers of MC steps, whereas BayesGE $1/k$ has the lowest error at high numbers of MC steps. On the 20 colour Potts model BayesGE $1/k$ outperforms all the other methods: it is only AIS that has a slightly lower RMSE at low numbers of MC steps. On the 100 colour Potts model both AIS and nested sampling performs best at low number of iterations, but at the end the two BayesGE methods have the lowest RMSE. If we exclude the AIS results on the Ising system the typical speed-up time to reach a prescribed accuracy between BayesGE $1/k$ and any of the other algorithms is of one order of magnitude, as illustrated in figure S3.

Here we compared the performance of the algorithms in terms of number of MC steps, as for most systems the actual simulation time will be dominated by the evaluation of the posterior target distribution, c.f. section 3.6. On the 20 colour 16×16 Potts model the running times for AIS and WL with 10^9 MC steps are approximately 3 minutes (C++ implementation), whereas the BayesGE methods uses approximately 6 minutes (simulation implemented in C++ and inference in Python). The time overhead for BayesGE can mainly be attributed to the posterior inference for the entropy. It has to be emphasised, however, that the Ising and Potts models have been specifically chosen due to the fact that the posterior can be evaluated in constant time, which is what makes the accurate comparison of the different inference algorithms tractable on a wide range of MC timescales. As a proof-of-principle we have also applied BayesGE and AIS to a binary restricted Boltzmann machine trained on the MNIST dataset (LeCun et al. 1998), with the dual purpose of demonstrating the application of our algorithm to a more computational intensive model having a semi-continuous energy spectrum and verifying that BayesGE approaches the same scaling behaviour as AIS in this case, see section S7.

5 CONCLUSION

In conclusion, we have presented a new method for enhancing sampling efficiency and estimating key multivariate integrals in high-dimensional probability models. The robustness and accuracy of the method has been demonstrated and compared with existing methodologies on two classical spin systems displaying cooperative transitions and multimodality.

We note that both the WL and the BayesGE algorithm in its present formulation require a reasonable binning to be known prior to the simulation. However, in contrast to the WL algorithm our update rule, equation (26), does not presume prior knowledge of the relevant energy range, which is a considerable advantage in systems where the ground state(s) or low temperature properties are unknown. Extensions of our method to online binning will be the subject of future work.

There are a number of other directions to be pursued in future works. First of all, we wish to pertain a more absolute interpretation of the covariance structure of the posterior, by ensuring that the likelihood function scales correctly with the number of *independent* observations. Secondly, it is of obvious interest to explore other prior functions and determine their proper domain of application. Thirdly, we aim to extend the method to larger systems as well as multivariate representations of the density function using approximate posterior inference. As our of method can be applied to any statistical model, we believe it should find important applications in a wide range of computational fields, including machine learning, statistics, statistical physics and bioinformatics, where inference in high-dimensional systems forms a central problem.

Acknowledgements

JF acknowledge funding from the Danish Council for Independent Research 0602-02909B. ZG acknowledge funding from EPSRC EP/I036575/1 and Google.

References

Beale, P. D. (1996). "Exact Distribution of Energies in the Two-Dimensional Ising Model". In: *Physical Review Letters* 76 (1), pp. 78–81.

Berg, B. A. and Neuhaus, T. (1992). "Multicanonical ensemble: A new approach to simulate first-order phase transitions". In: *Physical Review Letters* 68 (1), p. 9.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Cambridge: Springer Verlag.

Ferkinghoff-Borg, J. (2002). "Optimized Monte Carlo analysis for generalized ensembles". In: *The European Physical Journal B* 29 (3), pp. 481–484.

Ferkinghoff-Borg, J. (2012). "Monte Carlo Methods for Inference in High-Dimensional Systems". In: *Bayesian Methods in Structural Bioinformatics*. Ed. by T. Hamelryck, K. Mardia, and J. Ferkinghoff-Borg. Statistics for Biology and Health. Springer Berlin Heidelberg, pp. 49–93.

Gelman, A. and Meng, X.-L. (1998). "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling". In: *Statistical Science* 13 (2), pp. 163–185.

Geyer, C. J. (1991). "Markov Chain Monte Carlo Maximum Likelihood". In: *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation of North America.

Hansmann, U. H. E. and Okamoto, Y. (1997). "Generalized-ensemble Monte Carlo method for systems with rough energy landscape". In: *Physical Review E* 56 (2), p. 2228.

Hastings, W. (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57 (1), p. 97.

Hesselbo, B. and Stinchcombe, R. B. (1995). "Monte Carlo Simulation and Global Optimization without Parameters". In: *Physical Review Letters* 74 (12), p. 2151.

Iba, Y. (2001). "Extended Ensemble Monte Carlo". In: *International Journal of Modern Physics C* 12 (5), p. 623.

Irbäck, A. and Potthast, F. (1995). "Studies of an off-lattice model for protein folding: sequence dependence and improved sampling at finite temperature". In: *Journal of Chemical Physics* 103, p. 10298.

Landau, D. P. and Binder, K. (2014). *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86 (11), pp. 2278–2324.

Lee, J. (1993). "New Monte Carlo algorithm: Entropic sampling". In: *Physical Review Letters* 71 (2), p. 211.

Lyubartsev, A. P., Martsinovski, A. A., Shevkunov, S. V., and Vorontsov-Velyaminov, P. N. (1992). "New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles". In: *The Journal of Chemical Physics* 96 (3), p. 1776.

Marinari, E. and Parisi, G. (1992). "Simulated Tempering: A New Monte Carlo Scheme". In: *Europhysics Letters* 19, p. 451.

Metropolis, N. and Ulam, S. (1949). "The Monte Carlo Method". In: *Journal of the American Statistical Association* 44 (247), pp. 335–341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21 (6), p. 1087.

Murray, I. (2007). "Advances in Markov chain Monte Carlo methods". PhD thesis. Gatsby computational neuroscience unit, University College London.

Murray, I., MacKay, D. J. C., Ghahramani, Z., and Skilling, J. (2005). "Nested sampling for Potts models". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. Platt, pp. 947–954.

Neal, R. M. (2001). "Annealed importance sampling". In: *Statistics and Computing* 11 (2), pp. 125–139.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). "A Unifying View of Sparse Approximate Gaussian Process Regression". In: *Journal of Machine Learning Research* 6, pp. 1939–1959.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Salakhutdinov, R. (2010). "Learning Deep Boltzmann Machines using Adaptive MCMC". In: *27th International Conference on Machine Learning*, pp. 943–950.

Skilling, J. (2006). "Nested sampling for general Bayesian computation". In: *Bayesian Analysis* 1 (4), pp. 833–859.

Swendsen, R. H. and Wang, J.-S. (1986). "Replica Monte Carlo Simulation of Spin-Glasses". In: *Physical Review Letters* 57 (21), pp. 2607–2609.

Wahba, G. (1978). "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression". In: *Journal of the Royal Statistical Society B* 40 (3), pp. 364–372.

Wang, F. and Landau, D. P. (2001). "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States". In: *Physical Review Letters* 86 (10), p. 2050.

Wu, F. Y. (1982). "The Potts model". In: *Reviews of Modern Physics* 54 (1), pp. 235–268.