Supplementary material for: 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

S1 The $1/k$ ensemble

As stated in section 2, the $1/k$ ensemble is constructed to put roughly equal probability on all value of the entropy. This can be understood by the following argument. For most systems the entropy is increasing fast with E. We can transform the $k(E) \equiv \int_{-\infty}^{E} g(E')\,\mathrm{d}E'$ to an integral over entropy and use that the entropy is growing fast with energy to make the approximation

$$\begin{aligned} k(E) &= \int_0^{s(E)} \left(\frac{dS}{dE}\right)^{-1} \exp(s)\,\mathrm{d}s \\ &\approx \left(\frac{dS(E)}{dE}\right)^{-1} \exp s(E)\,\Delta s\ , \end{aligned} \tag{S1}$$

where Δs is the appropriate scale of s. Transforming the distribution of $p(E|w)$, equation (6), to the distribution over the entropy

$$p(s|w) = \left(\frac{ds}{dE}\right)^{-1} \exp(w(E) + s(E)) \tag{S2}$$

and inserting the approximation derived into $w(E) = -\log k(E)$, we arrive at a flat distribution in entropy.

S2 The Wang–Landau algorithm

Arguably, the most successful GE learning algorithm having a similar domain of application as BayesGE is the *Wang–Landau* (WL) algorithm (Wang and Landau 2001). In this method, a flat histogram in $P(E|w)$ is enforced by increasing the entropy estimate $\hat{\mathbf{s}}$ by a constant modification factor $\log f > 0$ each time an energy bin is visited, using a predefined binning procedure. This modification procedure is repeated until the accumulated histogram satisfy some prescribed flatness criterion, which is periodically checked. At this point, the histogram is reset and a new iteration $t+1$ is started with a reduced parameter f according to the recipe $f_{t+1} = \sqrt{f_t}$. This means that $f_t \to 0$ as $t \to \infty$, and the weights $\mathbf{w} = -\hat{\mathbf{s}}$ will reproduce the multicanonical ensemble. Partition functions are then estimated using equation (8).

S3 The generalised multi-histogram equations

The maximum likelihood estimate (MLE) of the entropy $\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} P(N|W,\mathbf{s})$, c.f. equation (17), can be found using the *generalised multi-histogram* (GMH) equations (S3) and (S4) (Ferkinghoff-Borg 2002). Expressed in terms of the density of states $\hat{\mathbf{g}} = \exp\hat{\mathbf{s}}$, we have

$$\hat{g}_j = \frac{\sum_{\tau=1}^{t} n_j^{(\tau)}}{\sum_{\tau=1}^{t} \chi_j^{(\tau)} \nu^{(\tau)} Z_{\mathbf{w}^{(\tau)}}^{-1} \exp w_j^{(\tau)}}\ , \tag{S3}$$

where $\chi_j^{(\tau)} = 1$ if $j \in \tilde{\mathcal{S}}^{(\tau)}$ and otherwise $\chi_j^{(\tau)} = 0$. $\tilde{\mathcal{S}}^{(\tau)}$ and $\nu^{(\tau)}$ are defined in section 3.2. The partition functions $\{Z_{\mathbf{w}^{(\tau)}}\}_{\tau=1}^{t}$ must be estimated self-consistently from equation (S3). This can be formulated as the roots of t nonlinear equations

$$\left\{ \sum_{j \in \tilde{\mathcal{S}}^{(\tau')}} \frac{\sum_{\tau=1}^{t} n_j^{(\tau)}}{\sum_{\tau=1}^{t} \chi_j^{(\tau)} \nu^{(\tau)} \frac{Z_{\mathbf{w}^{(\tau')}}}{Z_{\mathbf{w}^{(\tau)}}} \exp(w_j^{(\tau)} - w_j^{(\tau')})} - 1 = 0 \right\}_{\tau'=1}^{t}\ . \tag{S4}$$

These can be solved effectively w.r.t. the t unknowns $\{Z_{\mathbf{w}^{(\tau)}}\}_{\tau=1}^{t}$ using the iterative Newton–Raphson method (Ferkinghoff-Borg 2012). Inserting the solution of these equations into equation (S3) leads to an estimate of the density of states.

Variants of these equations are also known as the *multihistogram* or *WHAM-equations* used for calculating respectively entropies of free energies for thermal ensembles, see Ferkinghoff-Borg (2012) for further details.

S4 Details on posterior inference

The prior and likelihood function only depends on $\mathbf{s}$ up to an additive constant. The translation invariance of

the likelihood function reflects the fact partition functions can only be estimated in a relative sense in the MCMC procedure, as discussed in section 2. We circumvent this invariance by conditioning the likelihood on $s_o = 0$ at a point E_o, where the entropy is large and thus easy to generate states from. It is straight forward to subsequently renomalise the results according to a known Z_0 (e.g. $Z_0 = 1$).

S5 Derivation of the closed form expression for setting the weights

To derive the closed form expression for the weights in (25), we expand the argument on the rhs. of equation (24) in terms of $\delta\mathbf{s} = \mathbf{s} - \bar{\mathbf{s}}$. To this end, let $z_j = z_j(w_j) = \exp(w_j + \bar{s}_j)$. Without loss of generality we may assume that $Z_{\mathbf{w}}(\bar{\mathbf{s}}) = \sum_{j=1}^{J} \exp(w_j + \bar{s}_j) = \sum_{j=1}^{J} z_j = 1$, so z_j is of the order of $1/L$ where $L = |\tilde{\mathcal{S}}|$ is the number of observed energies. To first order in $\delta\mathbf{s}$ we then have

$$\begin{aligned} -\log\left(Z_{\mathbf{w}}(\delta\mathbf{s})\right) &= -\log\left(\sum_{j=1}^{J} \exp(z_j + \delta s_j)\right) \\ &\approx -\mathbf{z}\cdot\delta\mathbf{s} \end{aligned} \tag{S5}$$

and

$$\begin{aligned} &\langle P(E_j|\mathbf{w})\rangle_{P(\mathbf{s}|N,W)} \\ &\quad\approx \langle \exp\left(\log(z_j) + \delta s_j - \mathbf{z}\cdot\delta\mathbf{s}\right)\rangle_{P(\mathbf{s}|N,W)} \\ &\quad= z_j \exp\left(\frac{1}{2}V_{jj} + \frac{1}{2}\mathbf{z}^{\intercal}\cdot V\cdot\mathbf{z}\right), \end{aligned} \tag{S6}$$

where V is the posterior covariance matrix. Since V is diagonally dominant and $z \sim \mathcal{O}(\frac{1}{L})$, the exponent is typically dominated by the first term. Consequently, we see that for both the multicanonical and $1/k$ ensemble we obtain the desired expectation values of $P(E_j|\mathbf{w}^{(t+1)})$ under the posterior, by simply setting

$$\tilde{\mathbf{w}}^{(t+1)} = \bar{\mathbf{w}}^{(t+1)} - \frac{1}{2}\operatorname{diag}(V)\ , \tag{S7}$$

where $\bar{\mathbf{w}}^{(t+1)}$ are the weights defined in equations (22) to (23).

S6 Simulation details

For all BayesGE simulations we used the default initial simulation time $\nu^{(1)} = 5000$ and increasment factor $\gamma = 2^{1/10}$. The weights was updated using the uncertainty approximation from equation (25). For BayesGE with multicanonical weights we used the index set $\tilde{\mathcal{S}}$ restricted to the observed support when calculating the MLE $\hat{\mathbf{s}}$ and the negative Hessian H. For BayesGE with $1/k$ weights we used we used the full index set $\mathcal{S}$ when calculating $\hat{\mathbf{s}}$ and the restricted index set $\tilde{\mathcal{S}}$ for when calculating H.

In AIS we used a linear cooling schedule for β and performed a cooling step in each Monte Carlo step. For AIS, simulations was run at different lengths in batches of 50 simulations. The WL algorithm was run using the modification factor $\log f_t = 2^{-t}$ for the t'th level of random walk and a flatness criterion of 80%. For nested sampling we used MCMC to sample from the likelihood constrained prior, and the plots were produced by increasing the number of particles. The number of MCMC step used to draw a sample from the constrained prior was keep fixed for each model and selected optimal by trail-and-error, see figure S5. For each number of particles, nested sampling was run for $2\tilde{N}\tilde{H}$ steps, where $\tilde{N}$ is the number of particles and $\tilde{H}$ is the logarithm of the fraction of prior mass that count the bulk of the posterior mass (Skilling 2006). $\tilde{H}$ was estimated by an initial set of simulations using the maximal number of particles for the given number of MCMC steps.

S7 Estimation of the partition function of a binary restricted Boltzmann machine

As a proof-of-principle we have compared the performance of BayesGE and AIS with respect to estimating the partition function of a binary restricted Boltzmann machine (RBM). The purpose of these simulations is to demonstrate the application of BayesGE to more computationally intensive models involving semi-continuous energies and to verify the discussion in section 3.6 regarding the computational overhead of the algorithm.

The RBM consists of 784 visible units and 500 hidden units trained on the MNIST dataset (LeCun et al. 1998) using persistent contrastive divergence (Tieleman 2008). We estimate the partition function for $\beta = 1$, and for sake of consistency with Grosse et al. (2013), we have chosen the natural counting measure for p_0, equation (1), implying that that $Z_0 = 2^{784+500}$. We used the model PCD(500) trained by Grosse et al. (2013) and for which accurate estimates of the partition function have been obtained based on extensive AIS simulations. We have used the estimate with the largest effective sample size, $\log Z_{\beta=1} = 418.26$, as the reference for the comparison of the two methods.

We ran BayesGE with $1/k$ weights using standard settings and divided the energy range $[-500; 5000]$ uniformly into 1024 bins. As for the spins systems we used a linear cooling schedule for β in AIS. For both

algorithms we use a single bit flip Metropolis proposal.

Figure S6 shows $\log Z_{\beta=1}$ and the RMSE of $\log Z_{\beta=1}$ as a function of MC steps. These results indicate that BayesGE performs comparable to AIS in terms of convergence and that the binning in a semi-continuous model does not pose any principle problems to our method. After 10^9 MC steps BayesGE is seen to otain a better estimate of $\log Z_{\beta=1}$ on average but has a higher variance than AIS. Here two observations are of particular interest. First, the BayesGE average estimate displays a plateau from $\sim 1{\cdot}10^8$ to $\sim 4{\cdot}10^8$ MC step which is also mirrored in the RMSE. Secondly, on average the AIS estimate seems to converge slower to the reference value than the BayesGE at longer simulation times. We believe both observations are related to the inherent difficulties in sampling across the cooperative transition from unordered images to digits in the RBM model using the single flip Monte Carlo simulation setup. In particular the latter observation tentatively demonstrates the merits of BayesGE compared to tempering methods with respect to such type of transition. We aim to study these aspects in more details in a forthcoming publication.

For both methods we measured the wall-clock simulation time up to 10^8 MC steps. The simulations were performed on a 2.5GHz AMD Opteron 6380 processor and the BayesGE algorithm was allowed to use multiple threads for the entropy inference step. Figure S7 shows the wall-clock simulation time as a function of the number of MC steps. As expected, AIS scales linearly with time and BayesGE approaches the same scaling behaviour within a small number of MC steps compared to the total number of steps required for convergence, c.f. section 3.6. After 10^8 steps the overhead of BayesGE entropy inference step is less than 20% compared to AIS.

References

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

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.

Grosse, R., Maddison, C. J., and Salakhutdinov, R. (2013). "Annealing between distributions by averaging moments". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, pp. 2769–2777.

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.

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

Tieleman, T. (2008). "Training restricted Boltzmann machines using approximations to the likelihood gradient". In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 1064–1071.

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.

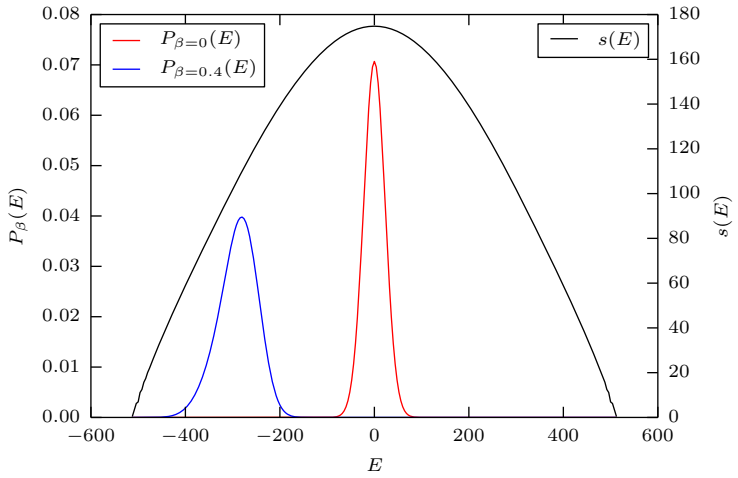

Figure S1: The black line shows the analytically calculated (Beale 1996) entropy function $s(E)$ for the 2D Ising model of size 16×16. The red curve shows the marginal distribution of the energy under a flat prior $P_0 = P_{\beta=0}$. The blue curve shows the marginal distribution over the energy induced by $P_{\beta=0.4}(\mathbf{x})$. Note that these marginal distributions can be calculated directly using equation (6) and $w_\beta(E) = -\beta E$.

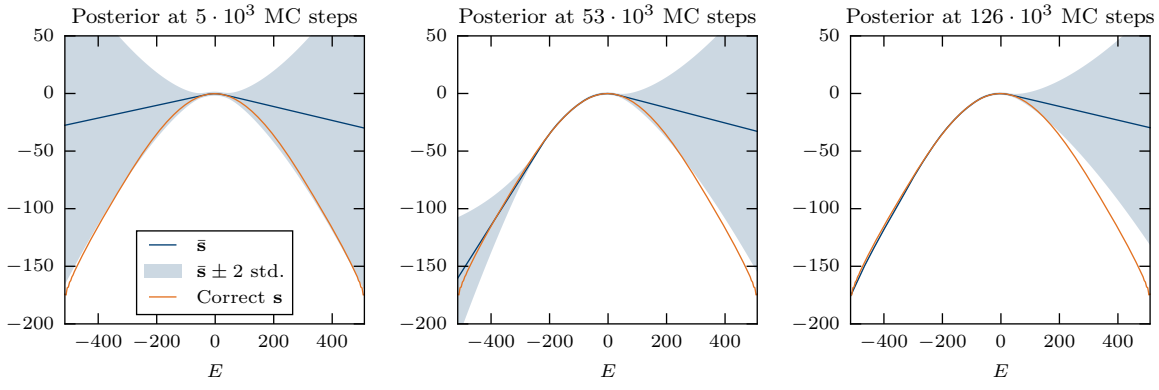

Figure S2: Examples of the posterior distribution of the entropy function $\mathbf{s}$ for BayesGE with $1/k$ weight at three different simulations lengths (corresponding to 1, 10 and 20 histograms) for the 2D Ising model of size 16×16. The orange line is the ground truth and the blue line is the posterior mean estimate $\bar{\mathbf{s}}$ with the shaded area showing $\pm$ two standard deviations.

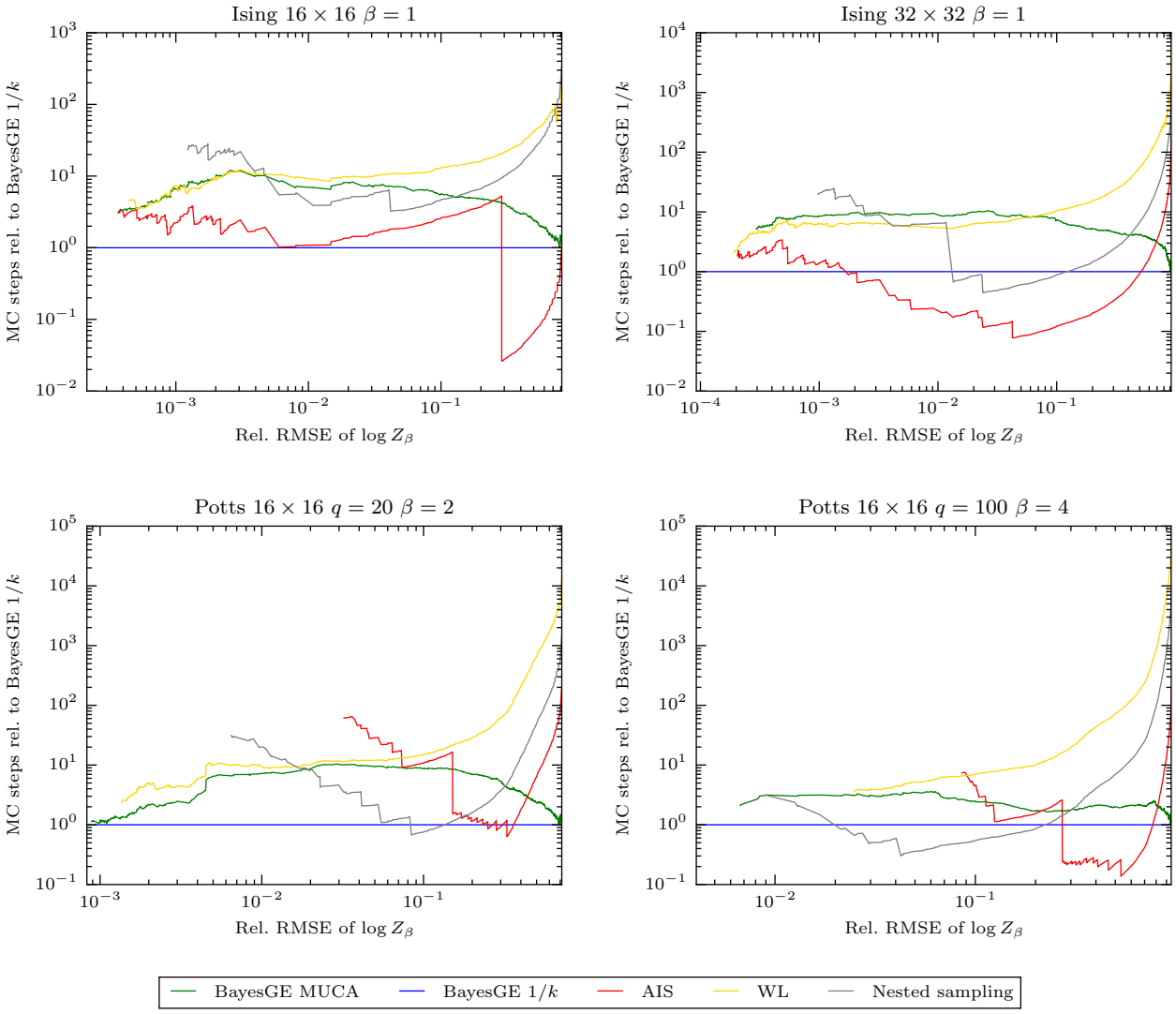

Figure S3: The plots show the relative number of MC steps need to obtain a given relative RMSE on the 2D Ising and Potts models with different sizes, number of colours (q) and values of β. The number of MC is measured relative to how many steps BayesGE $1/k$ uses to obtain the same relative RMSE. A relative number of MC above 1 means that a given method uses more MC steps than BayesGE $1/k$ to obtain a given error. A relative number of MC steps below 1 means that the method uses fewer MC steps.

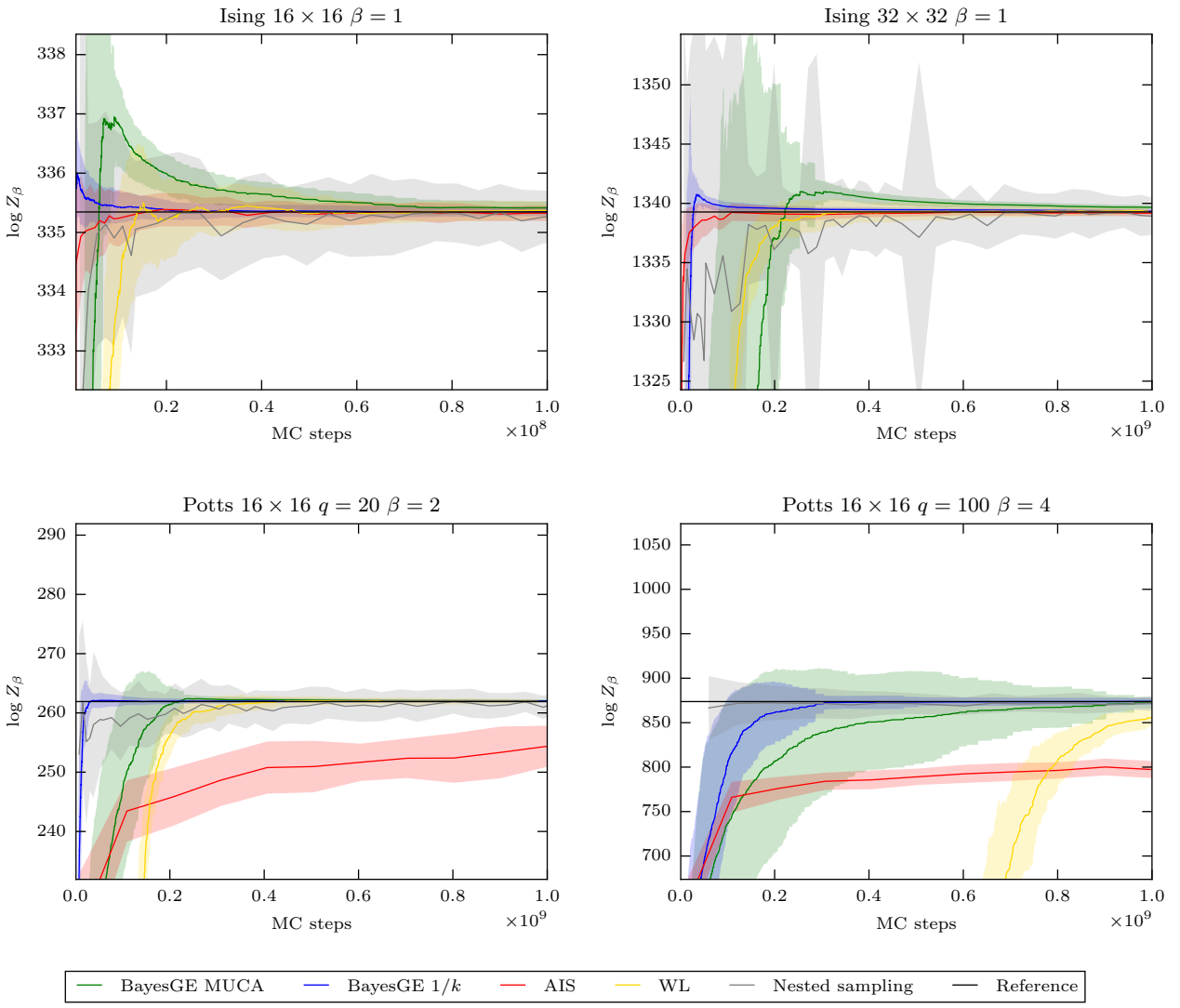

Figure S4: The log partition function $\log Z_\beta$ as a function of the number Monte Carlo (MC) steps for simulations on 2D Ising and Potts models with different sizes, number of colours (q) and values of β. For each method and each model, the line shows the average value of $\log Z_\beta$ over 50 independent simulations, and the shaded area shows $\pm$ one standard deviations. The black line shows the reference value for $\log Z_\beta$. For the Ising models, the reference value was calculated analytically (Beale 1996). For the Potts models the reference value was calculated as the average $\log Z_\beta$ over 50 independent WL simulations using 10^{10} MC steps.

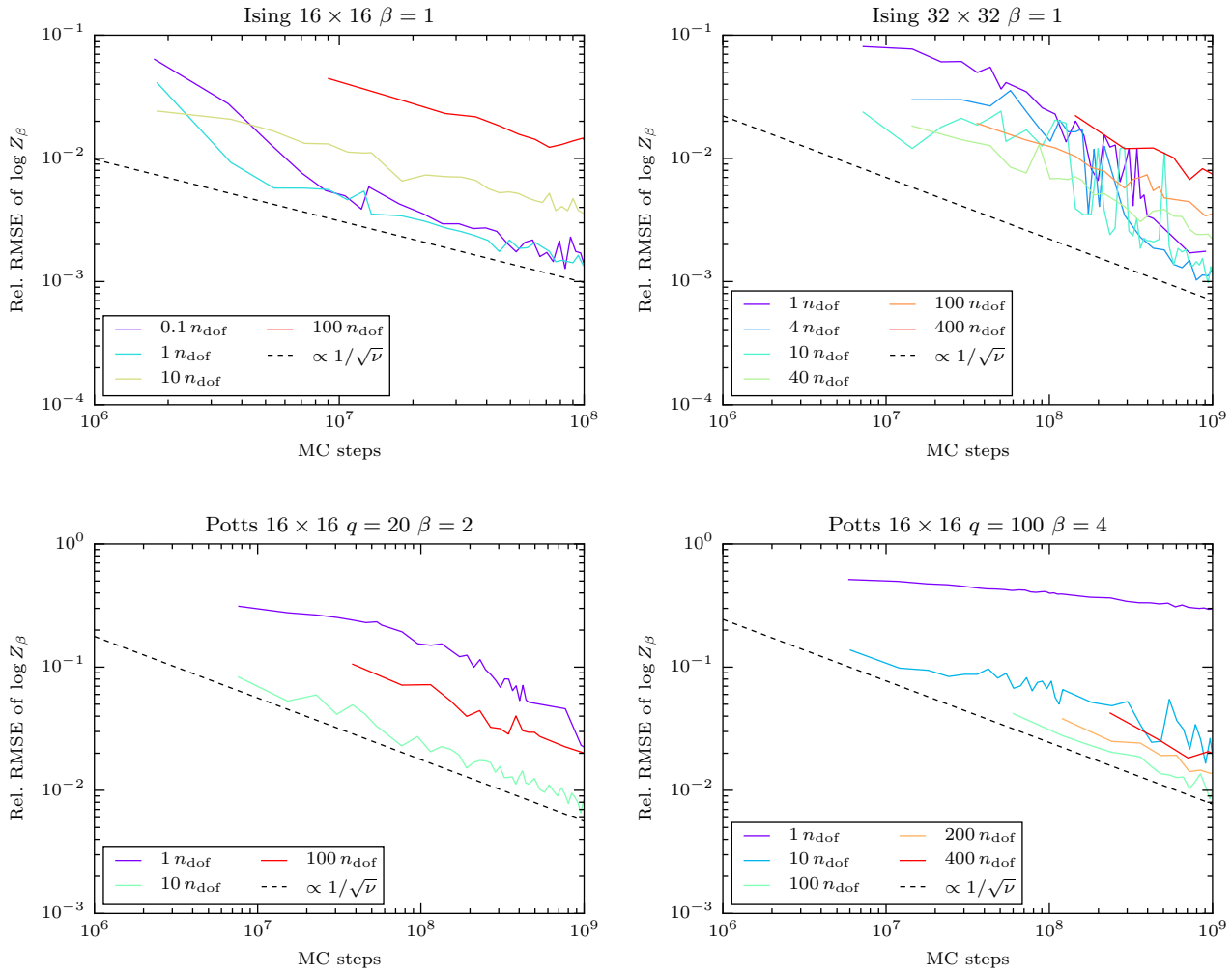

Figure S5: Illustration of the trial-and-error scheme used for selecting the number of MCMC steps used in *nested sampling* to draw *one* sample from the likelihood constrained prior. The plots shows the relative root mean square error (RMSE) of $\log Z_\beta$ (over 50 independent simulations) as a function 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 each model we tested a number of different values for the number MCMC step used to draw *one* sample from the constrained prior. The number of MCMC steps are expressed in terms of a factor times the *number of degrees of freedom* $n_{\text{dof}} = L^2$ of the model. For each model, the optimal number of MCMC steps was selected base on the (subjectively) optimal curve.

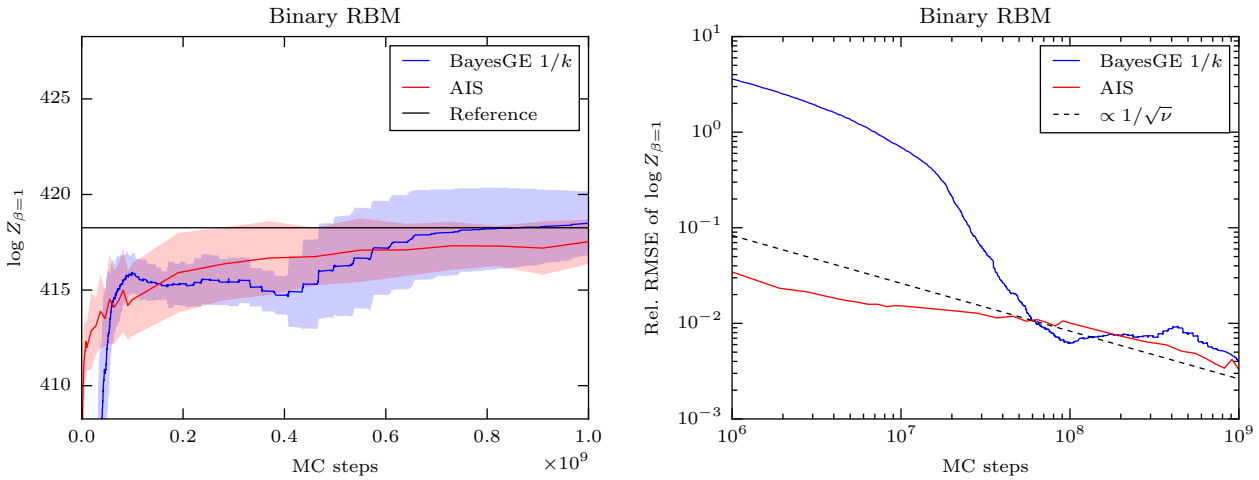

Figure S6: Simulation results for the binary RBM PCD(500) (Grosse et al. 2013) with 784 visible units and 500 hidden units trained on the MNIST dataset. The left plot shows $\log Z_{\beta=1}$ as a function of the number Monte Carlo (MC) steps and the right plot shows the RMSE of $\log Z_{\beta=1}$ as a function of MC steps. For each algorithm the results are averaged over 50 independent simulations and the shaded area on the left plot shows $\pm$ one standard deviation. The reference value $\log Z_{\beta=1} = 418.26$ is obtained from Grosse et al. (2013).

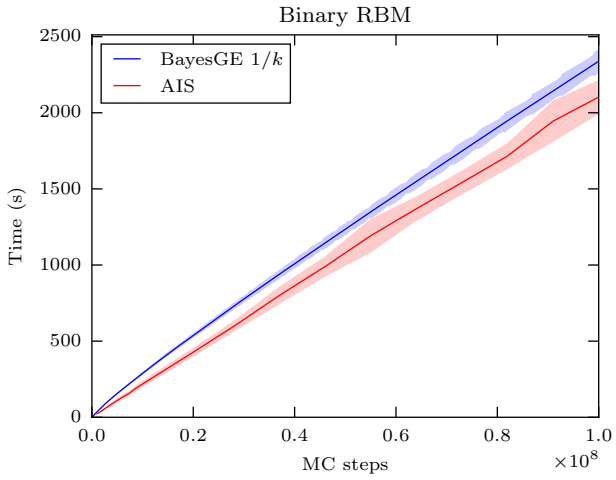

Figure S7: The wall-clock simulation time (measured in seconds) as a function of the number Monte Carlo (MC) steps for the binary RBM PCD(500) (Grosse et al. 2013) with 784 visible units and 500 hidden units trained on the MNIST dataset.