Nested sampling with demons

Michael Habeck

Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen, Germany
Felix Bernstein Institute for Mathematical Statistics in the Biosciences, University of Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany

Abstract. This article looks at Skilling's nested sampling from a physical perspective and interprets it as a microcanonical demon algorithm. Using key quantities of statistical physics we investigate the performance of nested sampling on complex systems such as Ising, Potts and protein models. We show that releasing multiple demons helps to smooth the truncated prior and eases sampling from it because the demons keep the particle off the constraint boundary. For continuous systems it is straightforward to extend this approach and formulate a phase space version of nested sampling that benefits from correlated explorations guided by Hamiltonian dynamics.

Keywords: Bayesian computation; Nested sampling; Monte Carlo simulation; Microcanonical ensemble; Demon
PACS: 02.50.Tt, 02.70.Rr, 02.70.Uu, 05.10.Ln

NESTED SAMPLING

Nested sampling [1] aims to compute the evidence

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

and, as a by-product, to draw samples from the posterior distribution

$$p(\theta) = \frac{1}{Z} L(\theta)\,\pi(\theta) \tag{2}$$

where θ is a d-dimensional parameter vector, $\pi(\theta)$ the prior probability and $L(\theta)$ the likelihood. If we succeed in solving tasks (1) and (2) we can carry out a complete Bayesian analysis including parameter estimation and model comparison.

In a nutshell, nested sampling builds on two observations. First, the d-dimensional evidence integral (1) can be reduced to a one-dimensional integral

$$Z = \int_0^1 L(X)\,\mathrm{d}X \tag{3}$$

that sums over likelihood weighted by the fraction of prior mass it encloses

$$X(\lambda) = \int_{L(\theta)\geq\lambda} \pi(\theta)\,\mathrm{d}\theta. \tag{4}$$

$L(X)$ is the inverse of $X(L)$ and different from the original likelihood function $L(\theta)$.

Second, if N particles $\theta_1, \ldots, \theta_N$ explore the prior above likelihood contour λ:

$$\theta_n \sim p(\theta|\lambda) = \frac{\Theta[L(\theta) - \lambda]}{X(\lambda)} \pi(\theta) \quad \text{where} \quad \Theta(x) = \begin{cases} 0; & x < 0 \\ 1; & x \geq 0 \end{cases}, \tag{5}$$

the associated prior masses $X_n \equiv X(L_n)$ enclosed by the contours $L_n \equiv L(\theta_n)$ follow a uniform distribution where $p(\theta|\lambda)$ denotes the prior subject to a likelihood constraint λ. The uniformity of the distribution of X_n is a consequence of definition (4) and can be viewed as a generalization of the probability transform. Because $X(\lambda)$ measures accumulated density, the unknown X_n can be ordered by sorting the states according to their likelihood. If the states are numbered such that $L_1 \leq L_2 \leq \cdots \leq L_N$, the prior masses will follow the reverse order: $X_1 \geq X_2 \geq \cdots X_N$. The fractional prior mass $X_n/X(\lambda)$ follows Beta$(N+1-n, n)$; notably the sampling distribution of the mass associated with the worst state is

$$X_1 \sim N \frac{X^{N-1}}{X(\lambda)^N}, \quad 0 \leq X \leq X(\lambda). \tag{6}$$

We can therefore *predict* the prior mass enclosed by the worst state.

Nested sampling proceeds stepwise starting with N states sampled from the unbounded prior ($\lambda = 0$). In each iteration k, the worst state defines a new likelihood contour λ_{k+1} that is used to restrict the prior in the next step. The prior mass associated with contour λ_k is estimated using order statistics (6); the initial mass being $X(0) = 1$. By construction, the survivors will already follow the truncated prior (5) at contour λ_{k+1}. The worst state is replaced by a new state that evolved from a randomly picked survivor.

A single nested sampling iteration moves from the current likelihood contour λ to the next $\lambda' > \lambda$. The overlap between two successive truncated priors $p(\theta|\lambda)$ and $p(\theta|\lambda')$ may be quantified by using the relative entropy [1]

$$H(\lambda \to \lambda') = \int p(\theta|\lambda') \ln[p(\theta|\lambda')/p(\theta|\lambda)] \, \mathrm{d}\theta = \ln[X(\lambda)/X(\lambda')]. \tag{7}$$

On average, the relative entropy changes by a constant amount controlled by the number of particles: $\langle H(\lambda \to \lambda') \rangle = -\langle \ln t \rangle_{t \sim \text{Beta}(N,1)} = 1/N$.

ANALOGY WITH STATISTICAL PHYSICS

In the following, we will interpret nested sampling as a method to solve computational problems in statistical physics [1, 2, 3, 4]. The parameters θ correspond to the microstate or configuration of a system whose potential energy $E(\theta) = -\ln L(\theta)$ is the negative log likelihood. The density of states (DOS) [5]

$$g(E) = \int \delta[E - E(\theta)] \, \pi(\theta) \, \mathrm{d}\theta \tag{8}$$

is the marginal distribution of log likelihood values over the prior. Likelihood contour λ corresponds to the energy limit $\varepsilon = -\ln \lambda$; the prior mass enclosed by ε is the cumulative

distribution function of the DOS:

$$X(\varepsilon) = \int_{E(\theta)\leq\varepsilon} \pi(\theta)\,\mathrm{d}\theta = \int_{-\infty}^{\varepsilon} g(E)\,\mathrm{d}E \tag{9}$$

where $X(\cdot)$ is now understood as a function of the energy rather than likelihood. The evidence integral (3) reduces to an evaluation of the partition function

$$Z(\beta) = \int e^{-\beta E(\theta)}\,\pi(\theta)\,\mathrm{d}\theta = \int e^{-\beta E}\,g(E)\,\mathrm{d}E \tag{10}$$

at inverse canonical temperature $\beta = 1$. This fact has driven the adaptation of thermal sampling methods such as simulated tempering [6] and parallel tempering [7] to a Bayesian context. Thermal sampling considers a series of canonical ensembles, $\pi(\theta)\exp\{-\beta E(\theta)\}$, at decreasing temperature and typically resorts to thermodynamic integration [8] to evaluate the partition function. In contrast, nested sampling aims to compute the evidence by estimating the DOS. Because it places the energy contours adaptively, nested sampling sidesteps a major problem with thermal methods which is to choose a good temperature schedule. Especially for systems undergoing a phase transition it is highly non-trivial to find well-balanced β-schedules.

There are other interesting quantities of statistical mechanics that can be related to nested sampling. The logarithm of the prior mass is proportional to Gibbs' definition of the microcanonical entropy (volume entropy): $S_G(E) = \ln X(E)$, while $S_B(E) = \ln g(E)$ is the standard Boltzmann definition (surface entropy) [9]. The microcanonical temperature is $T_G(E) = (\partial_E S_G)^{-1} = X(E)/g(E) \geq 0$ [10, 11], i.e. the reciprocal temperature $\beta_G = 1/T_G = \partial_E S_G$ tracks the speed at which the volume of the space of accessible configurations compresses. The compression achieved by a single iteration amounts to the difference in volume entropy: $H(\varepsilon' \to \varepsilon) = S_G(\varepsilon') - S_G(\varepsilon) = \int_{\varepsilon}^{\varepsilon'} \beta_G(E)\,\mathrm{d}E$, thus β_G measures the entropy production when moving in the reverse direction $\varepsilon \to \varepsilon'$. The optimal cooling protocol entails a constant relative entropy, which is achieved when the decrement of successive energy bounds is $\varepsilon_k - \varepsilon_{k+1} \approx 1/N\beta_G(\varepsilon_k)$. Therefore, a histogram of all energy bounds will follow β_G.

In summary, nested sampling supports a microcanonical rather than a canonical view. Instead of the canonical temperature β it uses the maximum attainable energy ε as control parameter. This is convenient because the energy can be evaluated directly at each microstate θ, whereas the canonical temperature of thermal algorithms is an ensemble property. The sequence of energy bounds ε_k does not need to be prescribed (as in *microcanonical annealing* [12]) but is found adaptively in the course of a nested sampling run. Nested sampling progresses optimally, at constant thermodynamic speed, slicing off a constant fraction of volume entropy in each step.

ENTER THE DEMON

We will now set up a microcanonical ensemble that combines the system with a demon to obtain the truncated prior distribution (5). For a fixed total energy ε consider the

microcanonical ensemble [10]:

$$p(\theta, D|\varepsilon) = \frac{1}{X(\varepsilon)} \delta[\varepsilon - D - E(\theta)] \Theta(D) \pi(\theta) \tag{11}$$

where $D \geq 0$ is the energy of a *demon* [13], an additional degree of freedom or *auxiliary variable* in statistical language. The marginal distribution over the states obtained by integrating out the demon's energy is the constrained prior (5). Thus, if we draw (θ, D) samples from (11) we are generating configurations from the truncated prior. The sampling procedure has to conserve the total energy ε. The Creutz algorithm [13] tells us how to do this:

1. Given the current configuration θ with energy $E(\theta)$ and the demon's energy D, generate a candidate state θ' from $\pi(\theta)$ with energy $E(\theta')$.
2. Propose a new state for the demon $D' = D - \Delta E$ where $\Delta E = E(\theta') - E(\theta)$.
3. If $D' > 0$, accept the proposal; else reject.

States generated by the Creutz algorithm will have energies below the energy bound: $E(\theta) \leq \varepsilon$.

We can now implement nested sampling as a demon algorithm. The demon removes high energy particles and stores their energy to define the next energy bound ε_k. After the "hot" particle has been taken out, a new particle is injected that does not exceed the new energy limit. One possibility is to pick a state θ from the set of survivors. A new configuration is obtained by providing the demon with energy $D = \varepsilon_k - E(\theta)$ and running the Creutz algorithm. The energies of the demon follow the distribution

$$p(D|\varepsilon) = \int p(\theta, D|\varepsilon) \, \mathrm{d}\theta = \frac{g(\varepsilon - D)}{X(\varepsilon)} \approx \frac{g(\varepsilon)}{X(\varepsilon)} \exp\{-D/T_B(\varepsilon)\} \tag{12}$$

where $T_B(E) = (\partial_E S_B)^{-1}$ is the microcanonical temperature based on the surface entropy. For large systems, T_G and T_B are virtually identical [11]. Therefore, the demon can serve as a thermometer: $T_G \approx \overline{D}$.

To illustrate these physical analogies, we apply nested sampling to the two-dimensional Ising model on an $L \times L$ lattice. The Ising model is a spin lattice model that recapitulates spontaneous magnetization phenomena characterized by a second order phase transition. The energy is $E(\theta) = \sum_{\langle i,j \rangle} \theta_i \theta_j$ where $\theta_i = \pm 1$ are the spin variables and the sum runs over nearest neighbors on two-dimensional regular grid. Figure 1 shows results for a system of size $L = 64$ (i.e. $d = 4096$) using $N = 100$ particles. For this particular run the estimated log evidence is $\ln Z = 5.3555 \times 10^3$, which comes very close to the exact value 5.355×10^3. Figure 1A shows the exact Gibbs entropy [14] and compares it with the estimate constructed by nested sampling. Nested sampling recovers $S_G(E)$ very accurately, which is reflected in the good estimate of the log evidence. Figure 1B shows the inverse temperature and confirms that a histogram of the energy bounds ε_k found by nested sampling indeed matches β_G. That is, nested sampling places the energy bounds such that the maximum attainable energy is steadily reduced, and the system is "cooled" in a controlled fashion. The cooling slows down at the phase transition, which is indicated by the peak of the heat capacity $C = 1/\partial_E T_G$ at

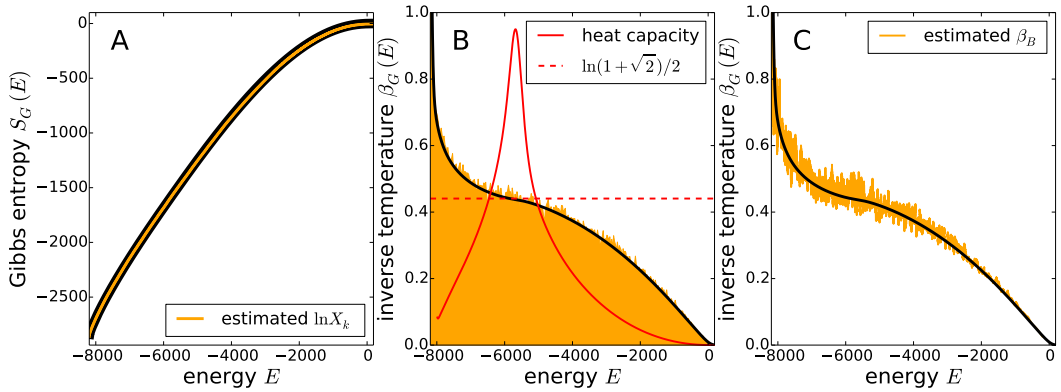

FIGURE 1. Results for the 64×64 Ising model. A: Exact Gibbs entropy (thick black curve), $\ln X(\varepsilon_k)$ estimated by nested sampling (orange line). B: Microcanonical inverse temperature β_G (black line). The orange area is a histogram over the $\sim 2.85 \times 10^5$ energy bounds ε_k found by nested sampling. The red curve indicates the microcanonical heat capacity (in arbitrary units so as to match the β_G range). C: The orange curve shows estimates of β_B obtained from the demon energies using a running average of window size 1000.

$E = -5680$. The peak corresponds to a canonical temperature of $\beta = 0.437$ (obtained upon numerical inversion of $\langle E \rangle_\beta = E$), which is close to the critical temperature in the thermodynamic limit $\ln(1+\sqrt{2})/2 \approx 0.44$ [15]. Figure 1C illustrates that the demon energies may indeed serve as a noisy thermometer.

RELEASING MORE DEMONS

Let us now consider a system with two demons, D and K, such that the microcanonical ensemble will be

$$p(\theta, D, K|\varepsilon) = \frac{1}{Y(\varepsilon)} \delta[\varepsilon - D - K - E(\theta)]\, \Theta(D)\, f(K)\, \pi(\theta) \tag{13}$$

where $f(K)$ is the energy distribution of the second demon. The prior volume is

$$Y(\varepsilon) = \int \Theta(\varepsilon - H)\, (f \star g)(H)\, \mathrm{d}H \tag{14}$$

where $(f \star g)(H)$ denotes the convolution of the second demon's energy distribution and measures the DOS evaluated at the total energy $H = K + E$. Tracking Y instead of X by exploring contours $\varepsilon > H$ computes the evidence of the extended system

$$Z_H = \int e^{-H(Y)} \mathrm{d}Y = \int e^{-H} (f \star g)(H)\, \mathrm{d}H = Z_K\, Z_E$$

by virtue of the convolution theorem for Laplace transforms (here, subscripts indicate the macrostate or sub-system, i.e. Z_E denotes the evidence of interest). If we know the Laplace transform of the demon's energy distribution, $Z_K = \int e^{-K} f(K)\, \mathrm{d}K$, we obtain the evidence $Z_E = Z_H / Z_K$.

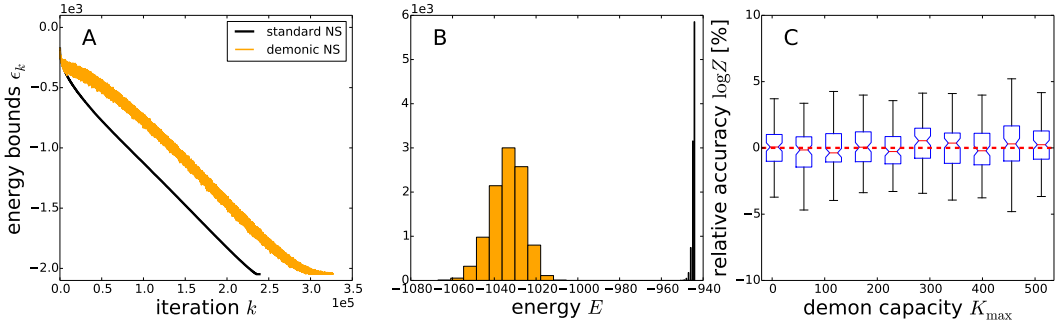

FIGURE 2. Demonic nested sampling of the ten state Potts model. A: Sampled energies of a 32×32 Potts model without (black) and with (orange) the additional demon. B: Energy samples at the first order phase transition ($\varepsilon = -944$) without and with additional demon shown as black and orange histograms, respectively. C: Benchmark on a 16×16 Potts model. Shown is the relative accuracy (in %) of the log evidence estimate for varying $K_{\max}$.

It is possible to obtain alternative versions of the microcanonical ensemble by coupling the system to an appropriate demon. For example, a d-dimensional harmonic oscillator demon with energy distribution $f(K) \propto \Theta(K_{\max} - K)\, K^{d/2-1}$ (with $K_{\max} > 0$ being the maximum energy the demon can absorb) results in the marginal distribution

$$p(\theta|\varepsilon) \quad \propto \quad \Theta[\varepsilon - E(\theta)]\, \pi(\theta) \times \begin{cases} [\varepsilon - E(\theta)]^{d/2}; & \varepsilon - E(\theta) \leq K_{\max} \\ K_{\max}^{d/2}; & \varepsilon - E(\theta) > K_{\max} \end{cases} \tag{15}$$

which is similar to the ensemble used in Ray's microcanonical Monte Carlo algorithm [16]. This ensemble differs from the truncated prior (5) by the additional factor $[\varepsilon - E(\theta)]^{d/2}$, which favors configurations with energies below the energy bound because the demon pushes the particle away from the constraint boundary. We can use Ray's Monte Carlo algorithm in the exploration phase and sample a new demon state afterwards by drawing from $p(K|\theta, \varepsilon)$. The system with highest total energy $H = E + K$ will then define the next energy bound ε. A drawback is that we have to deconvolve the DOS of the joint system $f \star g$ in order to recover g.

Tests of nested sampling with a second demon were run on the ten state Potts model [17] with a coupling constant of $J = 2$. We compare standard and demonic nested sampling of the 32×32 Potts model where a demon with capacity $K_{\max} = 2L^2$ was used. Demonic nested sampling took approximately one third longer than standard nested sampling. The estimated log evidences are comparable in accuracy with a relative error of 0.27% and 0.25% without and with demon, respectively. Figure 2A plots the energies of the sampled configurations. By construction the energies in the standard mode decrease monotonically, whereas in the demonic mode they fluctuate around a decaying average. The amount of scatter, and thereby also the length of the run, can be controlled by the demon's capacity $K_{\max}$. These fluctuations may help to explore configuration space more exhaustively (Fig. 2B). Systematic tests were run on the 16×16 Potts model. During these runs the demon's capacity was varied. For every $K_{\max}$, 100 repetitions where carried out using $N = 10$ particles. Figure 2C demonstrates that the accuracy of the estimated log evidence is not affected by the introduction of

the demon. Although there seems to be no advantage for the particular case of the Potts model, it may help to combine a system with multiple demons in other cases.

NESTED SAMPLING IN PHASE SPACE

Among its benefits is to use the demonic degrees of freedom during sampling in order to transfer energy between the main system and the demons. For continuous systems it is more natural set up the ensemble in phase space rather than configuration space. To do so, we unfold the demon and represent its energy distribution as the marginal distribution over microscopic demon variables ξ such that $f(K) = \int \delta[K - K(\xi)]\,\mathrm{d}\xi$. For every parameter θ_i we introduce associated momenta ξ_i resulting in $2d$ parameters in total. We set the demon's energy to the kinetic energy $K(\xi) = \frac{1}{2}\sum_{i=1}^{d} \xi_i^2$ whereupon the marginal distribution over configuration space is given by Eq. (15).

As in Hybrid Monte Carlo (HMC) [18], the dynamics defined by the Hamiltonian $H(\theta,\xi) = K(\xi) + E(\theta)$ is a useful guide in the exploration phase. Furthermore, it is convenient to augment the positions and momenta by one dimension to implement the demon D as a one-dimensional harmonic oscillator such that $D = (\xi_{d+1}^2 + \theta_{d+1}^2)/2$. The following microcanonical algorithm explores phase space under a constraint on the Hamiltonian ($\varepsilon > H$):

1. Given the current configuration θ with energy $E(\theta)$, set $\theta_{d+1} = 0$. Generate momenta from a $(d+1)$-dimensional standard Normal distribution, $\xi \sim N(0,1)$, and scale them such that the overall kinetic energy matches the available excess energy $K(\xi) + D = \varepsilon - E(\theta)$.
2. Simulate Hamiltonian dynamics in $2(d+1)$-dimensional phase space using the leapfrog algorithm to propose positions θ' and momenta ξ' with total energy $H' = E(\theta') + K(\xi')$ and $D' \geq 0$.
3. Accept if $H' < \varepsilon$, else reject.

Because the leapfrog algorithm approximately conserves the total Hamiltonian $H + D$ and $D \geq 0$, the proposal has a good chance to be accepted. In contrast to Constrained HMC [19] or Galilean Monte Carlo [20], the particle experiences forces from the likelihood during the dynamics and not only when it bumps into the constraint boundary whereupon it is reflected. The presence of the demon softens the constraint boundary; the extent of the zone in which the particle feels the presence of the boundary is determined by the demon's capacity $K_{\max}$. Alternatively, we could apply standard HMC directly to the constrained prior (15).

To illustrate the last point, we applied nested sampling to a small protein system, the 20-residue GS peptide, using 34 distances from Cavalli et al. [21]. GS peptide folds into a three stranded anti-parallel beta-sheet (Fig. 3A). The distance data were analysed using Inferential Structure Determination (ISD) [22]. A lognormal distribution serves as likelihood; a purely repulsive prior mimicks excluded volume effects; configurations were explored using HMC. Nested sampling with 100 particles and a demon efficiently locates minimum energy states (Fig. 3B) close to the ground state, as is indicated by the root mean square deviation (RMSD) to the native structure (Fig. 3C).

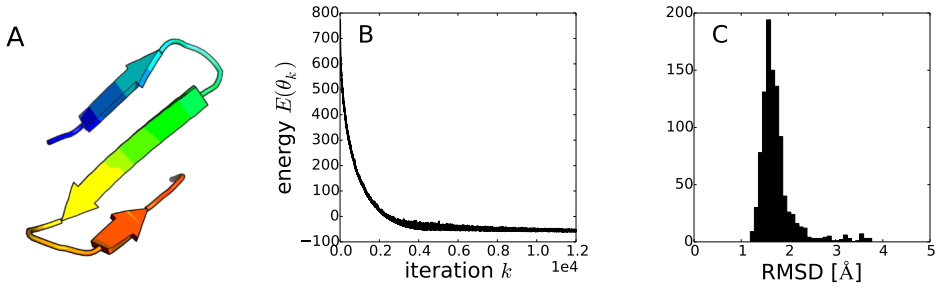

FIGURE 3. Nested sampling applied to GS peptide. A: Native structure of GS peptide. B: Evolution of the energy during nested sampling. C: Distribution of the structure's accuracy measured by the RMSD to the native structure.

CONCLUSION

Nested sampling is a powerful Monte Carlo method for Bayesian computation that has many desirable features also from a physical point of view. It constructs an adaptive cooling schedule such that the system moves at constant thermodynamic speed. A requirement is to draw configurations from the microcanonical ensemble, which may be achieved with the help of demons. Demons may also be used to sculpt the shape of the ensemble such that sampling of the extended system becomes easier.

ACKNOWLEDGMENTS

This work was supported by Deutsche Forschungsgemeinschaft (DFG) grant HA 5918/1-1.

REFERENCES

1. J. Skilling, *Bayesian Analysis* **1**, 833–860 (2006).
2. L. B. Partay, A. P. Bartok, and G. Csanyi, *J Phys Chem B* **114**, 10502–10512 (2010).
3. H. Do, J. D. Hirst, and R. J. Wheatley, *J Phys Chem B* **116**, 4535–4542 (2012).
4. S. O. Nielsen, *J Chem Phys* **139**, 124104 (2013).
5. A. I. Khinchin, *Mathematical Foundations of Statistical Mechanics*, Dover, 1960.
6. E. Marinari, and G. Parisi, *Europhys. Lett.* **19**, 451–458 (1992).
7. R. H. Swendsen, and J.-S. Wang, *Phys. Rev. Lett.* **57**, 2607–2609 (1986).
8. J. G. Kirkwood, *J. Chem. Phys.* **3**, 300–313 (1935).
9. M. Campisi, *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **36**, 275 – 290 (2005).
10. E. M. Pearson, T. Halicioglu, and W. A. Tiller, *Phys. Rev., A* **32**, 3030–3039 (1985).
11. J. Dunkel, and S. Hilbert, *Nat. Phys.* **10**, 67–72 (2014).
12. S. T. Barnard, Stereo matching by hierarchical, microcanonical annealing, Tech. rep., DTIC Document (1987).
13. M. Creutz, *Phys. Rev. Lett.* **50**, 1411–1414 (1983).
14. P. D. Beale, *Phys. Rev. Lett.* **76**, 78–81 (1996).
15. L. Onsager, *Phys. Rev.* **65**, 117–149 (1944).
16. J. R. Ray, *Phys. Rev., A* **44**, 4061–4064 (1991).

17. I. Murray, D. J. C. MacKay, Z. Ghahramani, and J. Skilling, "Nested sampling for Potts models," in *Advances in Neural Information Processing Systems 18*, edited by Y. Weiss, B. Schölkopf, and J. Platt, MIT Press, Cambridge, MA, 2006, pp. 947–954.
18. S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth, *Phys. Lett. B* **195**, 216–222 (1987).
19. M. Betancourt, *AIP Conference Proceedings* **1305**, 165–172 (2011).
20. J. Skilling, *AIP Conference Proceedings* **1443**, 145–156 (2012).
21. A. Cavalli, C. Camilloni, and M. Vendruscolo, *J Chem Phys* **138**, 094112 (2013).
22. W. Rieping, M. Habeck, and M. Nilges, *Science* **309**, 303–306 (2005).