

Stochastic Annealing

Robin C. Ball,^{1,*} Thomas M. A. Fink,^{2,3,†} and Neill E. Bowler^{1,‡}

¹*Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom*

²*Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75005 Paris, France*

³*Theory of Condensed Matter, Cavendish Laboratory, Cambridge CB3 0HE, United Kingdom*

(Received 14 January 2003; published 14 July 2003)

We show how to simulate a system in thermal equilibrium when the energy cannot be evaluated exactly: the error distribution needs to be symmetric, but it does not need to be known. We also solve the Ceperley-Dewing version of this problem, where the error distribution is taken to be fully known. These underlying ideas give an effective optimization strategy for problems where the evaluation of each design can be sampled only statistically, including an application to protein folding.

DOI: 10.1103/PhysRevLett.91.030201

PACS numbers: 02.60.Pn, 05.10.Ln, 02.50.Ng

We consider thermal equilibrium simulation of systems in which the energy of any given state is either not known exactly or else can much more cheaply be estimated. A classic example in physics is where each energy calculation itself involves sampling over a distribution. When the energy errors have known Gaussian distribution, an exact solution was noted by Ceperley and Dewing [1], and shown to be much more efficient than direct estimate of relative probability as used in lattice gauge theory (LGT) computations following Kennedy and Kuti [2]. An independent approximate attempt was also made by Krajčí and Hafner [3]. In this Letter we show how, by stochastic annealing, thermal equilibrium distributions can be sampled exactly and efficiently, even where the distribution of error is not exactly known.

These thermal sampling techniques can be applied to optimization problems where the objective function is analogously difficult to evaluate, using simulated annealing [4] (meaning simulated cooling) or related methods [5,6]. As an example we design model protein molecules to fold as fast as possible, where the only way to evaluate a particular design is to run a sample of folding simulations. Others have developed more empirical methods [7–9] for similar problems, but none of these is underpinned by simulation of true thermal equilibrium.

In a thermal equilibrium ensemble the probability of the system occupying a state μ with energy $E(\mu)$ is

$$P(\mu) \propto e^{-\beta E(\mu)}, \quad (1)$$

where $\beta = \frac{1}{T}$ is the reciprocal temperature. This distribution can be sampled by a Markov process, in which the system makes transitions (moves) from one state to another with rate constants $K(\mu \rightarrow \nu)$ which depend only on the two states concerned. Thermal equilibrium is achieved through detailed balance provided the move set is ergodic and the rate constants obey

$$\frac{K(\mu \rightarrow \nu)}{K(\nu \rightarrow \mu)} = e^{-\beta \Delta E}, \quad (2)$$

where ΔE is the energy difference $E(\nu) - E(\mu)$.

$K(\mu \rightarrow \nu)$ is typically the combination of an attempt frequency to move to state ν given that the system is in state μ , which we set to unity hereafter, multiplied by an acceptance probability. This acceptance probability must obey $0 \leq K \leq 1$. The Metropolis algorithm [10] is fully specified by requiring that $K(\Delta E) = 1$ for $\Delta E < C$, with maximal C (maximizing acceptance rates) giving

$$K_{\text{Metropolis}}(\Delta E) = \min(1, e^{-\beta \Delta E}), \quad (3)$$

whereas the Glauber acceptance function [11] arises by requiring that $K(\Delta E) + K(-\Delta E) = 1$, leading to

$$K_{\text{Glauber}}(\Delta E) = 1/(1 + e^{\beta \Delta E}). \quad (4)$$

These are compared graphically in Fig. 1.

We consider the case when the true energy change is not known exactly, and we must accept each move Γ with probability $A(x)$ based on $x = \Delta E(\Gamma) + y$ which is only an *estimate* of the true energy change $\Delta E(\Gamma)$. We assume the errors y are statistically independent with zero mean and distribution $f(y | \Gamma)$. The net probability of accepting a move whose true energy change is ΔE is then given by

$$K(\Delta E) = \int_{-\infty}^{\infty} f(y | \Gamma) A[\Delta E(\Gamma) + y] dy, \quad (5)$$

and it is our aim to choose $A(x)$ such that $K(\Delta E)$ satisfies detailed balance (2), the problem posed by Ref. [1].

We gain insight by considering the crude choice $A(x) = 1$ for $x < 0$ and $A(x) = 0$ otherwise. This simple strategy can give a good approximation to Eq. (2) of great value in optimization. The resulting acceptance function K when f is a Gaussian distribution is graphed in Fig. 1. The resemblance between this and the Glauber acceptance function (whose symmetry it shares) is striking, showing how the random energy errors make the selection look thermal—although in this case the match is not exact. The standard deviation σ of the Gaussian distribution controls an approximate effective temperature using this rule, as inverting Eq. (2) gives

$$T \simeq \sqrt{\frac{\pi}{8}} \sigma [1 - 0.018(\Delta E/T)^2 + \text{order}(\Delta E/T)^4]. \quad (6)$$

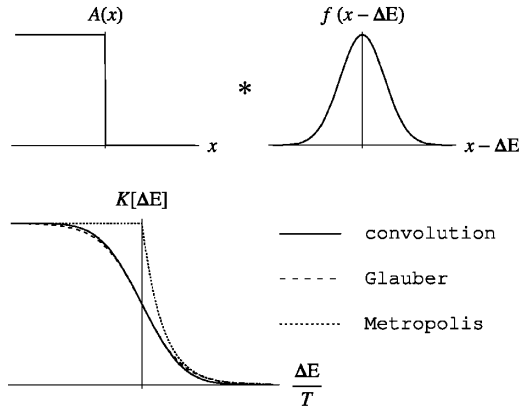


FIG. 1. A simple approximate stochastic annealing is obtained by accepting moves on the basis of the sign of the estimated energy change x . The acceptance probability as a function of the true underlying energy change ΔE is then given by a convolution of the true underlying energy change with the error distribution. As shown for the Gaussian distribution case, this gives an excellent approximation to the exact Glauber acceptance rule.

For many optimization purposes the small departure from detailed balance, due to the energy dependent terms, is not a problem.

Figure 2 shows how we successfully applied the above to optimize model protein folding: for each design change considered we obtained estimates of the change in mean folding time from a limited sample of folding simulations, and as the annealing proceeded we gradually cooled the system by increasing the sample sizes leading to reduced error size and reduced temperatures through Eq. (6). Our results show that optimizing folding performance selects sequences which fold with considerably less than maximal thermal stability, and the same appears to happen in nature: real (nonhomologous, two-state) proteins exhibit only marginal correlation between folding rate and stability [13]. In another paper [14] we show that our method can be exploited to unravel a benchmark problem in stochastic optimization, the probabilistic traveling salesman problem [15,16].

We now find $A(x)$ properly, such that K exactly satisfies detailed balance, Eq. (2), assuming for the present that the error distribution $f(y|\Gamma)$ is fully known.

Exact thermal sampling leads to some bounds on the behavior of $A(x)$ and of $f(y|\Gamma)$, in the latter case restricting when stochastic annealing is achievable. Detailed balance requires $\int_{-\infty}^{\infty} f(y|\Gamma)A[y + \Delta E(\Gamma)]dy = e^{-\beta\Delta E}K(-\Delta E) \leq e^{-\beta\Delta E}$ and for large positive ΔE this restricts all contributions to the left-hand side. First consider $y \approx 0$ for which $f(y|\Gamma)$ is not expected to be small. Then the exponential decay on this contribution must come from $A(x)$ [17] falling off at least as fast as $e^{-\beta x}$ as $x \rightarrow \infty$, which becomes important in our later analysis. Second, consider $y + \Delta E \approx 0$, for which $A(x)$ cannot become small or we would be heavily rejecting even moves which appear to be downwards in

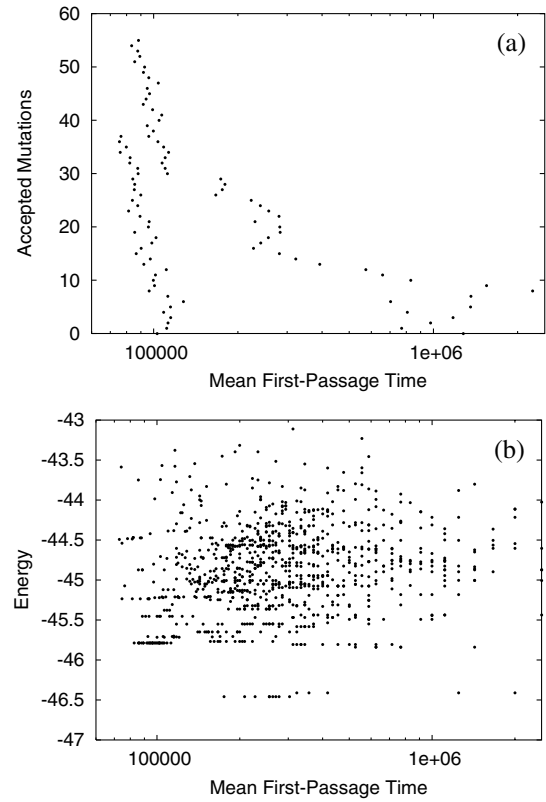


FIG. 2. Stochastic annealing applied to protein folding, in a simple cubic lattice model. The chains were 27 units long and the space of amino acid sequences was explored for the ability of the molecule to fold spontaneously to a fixed $3 \times 3 \times 3$ cubic target conformation. (a) The upper panel shows the result of two stochastic annealing simulations in which successive mutations of a starting sequence were accepted if a limited sampling of the mean first passage time (to the target) was improved. Increasing the depth of sampling as the annealing proceeded (y axis) provided the analog of lowering the temperature. (b) For comparison the lower panel shows the result of an extensive search over sequences guided by thermal stability (energy in the target conformation). Clearly individual runs of stochastic annealing find folding speeds approaching the fastest available, and much better than would be achieved by seeking minimum energy (the Shakhovich scheme [12]). Their energies (left -40.7 and right -38.5) are well above those selected by energy minimization.

energy. Then the exponential decay on this contribution must come from $f(-\Delta E|\Gamma)$ falling off at least as fast as $e^{-\beta\Delta E}$ for $\Delta E \rightarrow \infty$. This latter sets a fundamental restriction on stochastic annealing: the probability for large energy increases to be estimated as negative must fall off faster than a thermal Boltzmann factor.

Introducing the substitutions $f(y|\Gamma) = e^{(\beta/2)y}g(y, \Gamma)$ and $A(x) = e^{-\beta x/2}a(x)$, detailed balance is achieved by

$$a(x, \Gamma) = \int_{-\infty}^{\infty} h(s)g(x-s, -\Gamma)ds, \quad (7)$$

provided $h(s)$ and correspondingly $\tilde{h}(p)$ below are even.

Here $-\Gamma$ labels the transition inverse to Γ , and Eq. (7) appears to abuse our notation, in that it was earlier implied that the acceptance function $A(x)$ did not depend on hidden details of the move Γ . Our later discussion will focus on the case where the distribution of error is only assumed to be symmetric, so that $f(y | \Gamma) = f(y | -\Gamma)$ and hence also $g(y | \Gamma) = g(y | -\Gamma)$, and this is the only real restriction required for the analysis in between. However, for practicality and motivation we present it as relating to the case of invariant error distribution, where $g(y | \Gamma) = g(y)$ with any symmetric dependence on Γ suppressed from the notation.

We now aim to choose h to maximize the move acceptance rates, which are governed by $A(x)$. We follow the Metropolis methodology in choosing $A(x)$ to be identically 1 below some threshold, $x < C$. Then using the Fourier-Laplace transform defined by $\tilde{f}(p) = \int_{-\infty}^{\infty} e^{-px} f(x) dx$, the transform of a becomes

$$\tilde{a}(p) = \int_{-\infty}^C e^{\beta x/2} e^{-px} dx + \int_C^{\infty} b(x) e^{-px} dx, \quad (8)$$

where $b(x) = A(x)e^{\beta x/2}$ for $x \geq C$, $b(x) = 0$ for $x < C$. From Eq. (7) the transform of the new function $b(x)$ is

$$\tilde{b}(p) = \tilde{h}(p)\tilde{g}(p) - \frac{e^{[(\beta/2)-p]C}}{\frac{\beta}{2} - p}. \quad (9)$$

The exponential bound we established on the right tail of $A(x)$ [17] implies that $\tilde{b}(p)$ is bounded for $\text{Re} p > -\frac{\beta}{2}$, whereas the integrability of $f(y)$ together with the exponential bound on its left tail imply only that $\tilde{g}(p)$ is bounded for $-\frac{\beta}{2} \leq \text{Re} p < \frac{\beta}{2}$. Thus $\tilde{h}(p)$ must be chosen to cancel both any divergences of $\tilde{g}(p)$ in $\text{Re} p \geq \frac{\beta}{2}$ and the apparent pole at $p = \beta/2$. We first turn to the Wiener-Hopf method [18] to define

$$\tilde{g}(p) = \tilde{g}_L(p)\tilde{g}_R(p), \quad (10)$$

where $\tilde{g}_L(p)$ is bounded and nonzero for $\text{Re} p < \frac{\beta}{2}$, and $\tilde{g}_R(p)$ is bounded and nonzero for $\text{Re} p > \frac{\beta}{2}$; it also follows from the bounded window for $\tilde{g}(p)$ that $\tilde{g}_R(p)$ is bounded for the wider range $\text{Re} p > -\frac{\beta}{2}$. Then by choosing $\tilde{h}(p) = \frac{B}{[(\beta/2)^2 - p^2] \tilde{g}_L(p)\tilde{g}_L(-p)}$, we can ensure that $\tilde{h}(p)\tilde{g}(p) = \frac{B}{[(\beta/2)^2 - p^2]} \frac{\tilde{g}_R(p)}{\tilde{g}_L(-p)}$ is duly bounded for $\text{Re} p > -\frac{\beta}{2}$ except for the (desired) pole at $p = \frac{\beta}{2}$. Choosing the constant $B = \beta \frac{\tilde{g}_L(-\frac{\beta}{2})}{\tilde{g}_R(\frac{\beta}{2})}$ makes the residue of this pole cancel when we reassemble the expression (9) for $\tilde{b}(p)$, leading to the acceptance function given by

$$\tilde{a}(p) = \tilde{h}(p)\tilde{g}(p) = \frac{\beta}{[(\beta/2)^2 - p^2]} \frac{\tilde{g}_R(p)\tilde{g}_L(-\frac{\beta}{2})}{\tilde{g}_R(\frac{\beta}{2})\tilde{g}_L(-p)}. \quad (11)$$

C drops out of this optimal solution because it simply reflects the partition we introduced in Eq. (8).

As a simple example of our approach above, we consider $f(y) = \frac{1}{2} e^{-\gamma|y|}$. This leads to $\tilde{g}(p; \gamma) =$

$\frac{\gamma^2}{(\gamma - \frac{\beta}{2} - p)(\gamma + \frac{\beta}{2} + p)}$. The choice of \tilde{g}_L and \tilde{g}_R is trivial by inspection, giving $\tilde{a}(p) = \frac{\beta}{(\beta/2)^2 - p^2} \frac{\gamma + \beta}{\gamma} \frac{\gamma - \frac{\beta}{2} + p}{\gamma + \frac{\beta}{2} + p}$ and hence $A(x) = \min(1, \frac{\gamma^2 - \beta^2}{\gamma^2} e^{-\beta x} + \frac{\beta^2}{\gamma^2} e^{-(\gamma + \beta)x})$. This acceptance function is positive definite for $\beta \leq \gamma$, corresponding to the range of achievable stochastic annealing discussed earlier, and as $\gamma \rightarrow \infty$ it duly recovers the Metropolis method. We have analyzed other simple cases such as a rectangular error distribution, the superposition of two exponentials as above, and the multiple convolution of exponentials, all leading to results of equivalent properties.

The analysis of a Gaussian error distribution turns out to be slightly singular, in that its corresponding $\tilde{g}(p)$ has no obvious Wiener-Hopf factorization. However, we can approach it by considering the case where x is taken to be a sum of N independent exponential-distributed variables, the error distribution approaching Gaussian form as $N \rightarrow \infty$. In this case $\tilde{g}(p) = \tilde{g}(p; \gamma)^N$ with $\gamma = \sqrt{2N}/\sigma$, where σ is the standard deviation of x , leading to $\tilde{a}(p) = \frac{\beta}{(\beta/2)^2 - p^2} \left(\frac{\gamma + \beta}{\gamma} \frac{\gamma - \frac{\beta}{2} + p}{\gamma + \frac{\beta}{2} + p} \right)^N \rightarrow \frac{\beta}{(\beta/2)^2 - p^2} e^{\beta(p - \beta/2)\sigma^2/2}$ as $N \rightarrow \infty$ at fixed σ . The corresponding acceptance function by inverse transformation is then given by

$$A(x) = \min(1, e^{-\beta(x + \beta\sigma^2/2)}). \quad (12)$$

This is the one exact closed form solution previously known [1], and its singular status might be interpreted as the source of difficulty in perturbing about it, where Ceperley and Dewing obtained nonconvergent series [1].

The optimal acceptance rules found above all obey $0 \leq A(x) \leq 1$ required for a probability, but it can be objected that our optimal method gives no guarantee of this outcome. We have discovered a general but suboptimal solution for the acceptance function which does assure the requirement. The key idea is that, given the result (11), our requirements on the factorization of $\tilde{g}(p)$ can be relaxed to require only that $\tilde{g}_R(p)$ is bounded for $\text{Re} p > -\frac{\beta}{2}$ and that $\tilde{g}_L(p)$ is nonzero for $\text{Re} p < \frac{\beta}{2}$. Then assuming that $f(y) = 0$ for $y < C$, an acceptable factorization is given by $\tilde{g}_R(p) = \tilde{g}(p)e^{pC}$, $\tilde{g}_L(p) = e^{-pC}$ leading to $\tilde{a}(p) = \frac{\beta}{[(\beta/2)^2 - p^2]} \frac{\tilde{g}(p)}{\tilde{g}(\frac{\beta}{2})}$, at which point we can let $C \rightarrow -\infty$ so no significant new restriction has been imposed on f . The resulting acceptance function can be expressed as the convolution

$$A(x) = \frac{1}{\tilde{f}(\beta)} \int_{-\infty}^{\infty} K_{\text{Metropolis}}(x - y) e^{-\beta y} f(y) dy. \quad (13)$$

It can be verified by direct substitution that this obeys the detailed balance condition (2), is manifestly positive definite, and gives maximum acceptance 1 as $x \rightarrow -\infty$. This acceptance probability is lower than the optimal form where we have calculated that, the key difference being that an interval where $A = 1$ is not imposed.

We were very surprised to note that Eq. (13) even suggests a way to generate correct sampling when the

error distribution is *not* known, by using sampling of the error distribution to evaluate the integral in the acceptance rule (13). The key to doing this is first to double up the sampling of the energy change estimate, using $x = (x_1 + x_2)/2$ and then to use $y = (x_3 - x_4)/2$ as a sampling of the energy change error, where x_i are four statistically independent calculations of the same energy change. Evidently y validly samples the distribution of error in x provided only that the distribution of error in the x_i is symmetric, leading to

$$A(x_1, x_2, x_3, x_4) = \frac{1}{D} \min[e^{(-\beta/2)(x_1+x_2)}, e^{(-\beta/2)(x_3-x_4)}], \quad (14)$$

where D is a numerical divisor discussed below.

The above “universal” acceptance rule is remarkable in that it contains only two assumptions, the first being symmetry of the error distribution: positive and negative errors of the same magnitude must be equally likely for a given move. The second is that, although we do not need to know the shape of the distribution nor its dependence on the particular move Γ , we do need bounds on its outer tail so that we can safely set the divisor D to ensure $A < 1$ for an acceptance probability. The same sort of prefactor issue arises in the method [2] established in LGT computations, but the difference is that our setting of D is limited only by fluctuations in energy change estimate, whereas the prevailing LGT method has to set the equivalent of D to counter the full range of Boltzmann factor with the underlying energy change included.

If the errors in the underlying x_i have finite range, it is trivial to render the acceptance probabilities bounded by unity. A more practical example is where we model the tails of the distribution of the x_i by a Gaussian with standard error σ ; then the probability that $A > 1$ arises is bounded by $\varepsilon < (\beta\sigma/2\sqrt{\pi}\ln D) \exp[-(\ln D/\beta\sigma)^2]$ and the cost of making this extremely small turns out to be quite modest. The efficient way to do this is to make each of the x_i be itself the average over n samplings of the energy change, so that $\sigma(n) = \sigma n^{-1/2}$ which entails computational cost per accepted move proportional to $c(n) = nD$. Optimizing the latter with respect to the choice of n at fixed ε leads to $c_{\min} = (\frac{\varepsilon}{2}\beta\sigma)^2 \ln(\beta\sigma/2\sqrt{\pi}\varepsilon)$, excluding doubly logarithmic terms. Even when the original x_i are not Gaussian, under multiple sampling they approach this and similar economies obtain. One can take an empirical approach to the prefactor $1/D$, monitoring the acceptance probabilities and cumulating $A > 1$ into excess statistical weights: this approach requires multiple runs of the Markov chain and gives good sampling only if the cumulated weights stay close to unity.

We are confident our new methods of exactly thermal stochastic annealing will find significant direct application. One source of error with knowable distribution is

computational rounding, and it is intriguing that we can in principle correct for this in thermal sampling. The really remarkable result is our universal sampling rule (14) which requires bounds only on the tails of the error distribution, provided the distribution is symmetric. This includes the Gaussian error distribution where the variance can only be estimated [1]. We have also begun to investigate how using richer linear combinations in the exponents of Eq. (14) can accommodate nonsymmetric unknown distributions.

Underpinned by the exact results, our analysis provides a powerful new tool in stochastic optimization. Generally it will be sufficient and convenient to use the approximately thermal method we presented, of simply accepting moves which appear to be an improvement. Then all the benefits of simulated annealing are obtained by deliberately using crude estimates for each decision.

This research was supported by BP and EPSRC.

*Electronic addresses: r.c.ball@warwick.ac.uk

†Electronic addresses: fink@lps.ens.fr

‡Electronic addresses: Neill.Bowler@metoffice.com

Present address: Met Office, Maclean Building, Crowmarsh-Gifford, Oxfordshire OX10 8BB, UK.

- [1] D. M. Ceperley and M. Dewing, *J. Chem. Phys.* **110**, 9812 (1999).
- [2] A. D. Kennedy and J. Kuti, *Phys. Rev. Lett.* **54**, 2473 (1985).
- [3] M. Krajčí and J. Hafner, *Phys. Rev. Lett.* **74**, 5100 (1995).
- [4] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, *Science* **220**, 671 (1983).
- [5] A. Harju *et al.*, *Phys. Rev. Lett.* **79**, 1173 (1997).
- [6] E. J. Anderson and M. C. Ferris, *SIAM J. Optim.* **11**, 837 (2001).
- [7] H. Robbins and S. Munro, *Ann. Math. Stat.* **22**, 400 (1951).
- [8] W. B. Gong, Y. C. Ho, and W. Zhai, in *Proceedings of the 31st IEEE Conference on Decision and Control* (IEEE, Piscataway, NJ, 1992), pp. 795–802.
- [9] A. A. Bulgak and J. L. Sanders, in *Proceedings of the 1988 Winter Simulation Conference* (IEEE, Piscataway, NJ, 1988), pp. 684–690.
- [10] N. Metropolis *et al.*, *J. Chem. Phys.* **21**, 1087 (1953).
- [11] R. J. Glauber, *J. Math. Phys. (N.Y.)* **4**, 294 (1963).
- [12] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 1282 (1995).
- [13] K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, *Biochemistry* **39**, 11 177 (2000).
- [14] N. E. Bowler, T. M. Fink, and T. M. Fink, *Phys. Rev. E* (to be published).
- [15] P. Jaillet, *Oper. Res.* **36**, 929 (1988).
- [16] P. Jaillet, *Math. Oper. Res.* **18**, 51 (1993).
- [17] Strictly, we bound the weight of $A(x)$ under integration.
- [18] P. M. Morse and H. Feshbach, in *Methods of Theoretical Physics* (McGraw-Hill, London, 1953), pp. 978–980.