Author(s): Robin C. Ball, Thomas M. A. Fink and Neill E. Bowler
Article Title: Stochastic Annealing
Year of publication: 2003
Link to published article:
http://dx.doi.org/10.1103/PhysRevLett.91.030201

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, England*
[2] *CNRS UMR 144, Institut Curie, 75005 Paris, France*
[3] *Theory of Condensed Matter, Cavendish Laboratory, Cambridge CB3 0HE, England*

(Dated: February 2, 2008)

We demonstrate that is it possible to simulate a system in thermal equilibrium even when the energy cannot be evaluated exactly, provided the error distribution is known. This leads to an effective optimisation strategy for problems where the evaluation of each design can only be sampled statistically.

PACS numbers: 02.60.Pn, 05.10.Ln, 02.50.Ng
Keywords: stochastic optimisation, simulated annealing

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 which occurs in Physics is where each energy calculation itself involves sampling over a distribution or numerical integration, or the estimation of parameters of a numerical model. In this letter we show how, by Stochastic Annealing, thermal equilibrium distributions can nevertheless be sampled exactly, the essence of our method being that the energy errors can be precisely absorbed as a contribution to thermal noise.

Our thermal sampling technique can be applied to optimisation problems where the objective function is analogously difficult to evaluate, using Simulated Annealing[1] (meaning simulated cooling) or related methods [2, 3]. As an example we consider designing 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. Confronted by similar problems others have developed more empirical methods [4, 5, 6], but none of these is underpinned by simulation of true thermal equilibrium.

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

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

where $\beta = \frac{1}{T}$ is the inverse temperature. It is convenient to sample this distribution by a Markov process, in which the system is allowed to make a transition (move) from one state to another with rate constant $K(\mu \to \nu)$ which depends only on the two states concerned. This is generally more efficient than trying to choose the states directly, provided we can assume that the move set is ergodic, meaning that all states can be reached (eventually) from any given starting state. The more strict condition of detailed balance,

$$P(\mu)K(\mu \to \nu) = P(\nu)K(\nu \to \mu), \tag{2}$$

imposes the correct equilibrium distribution provided

$$\frac{K(\mu \to \nu)}{K(\nu \to \mu)} = e^{-\beta\,\Delta E} \tag{3}$$

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

Whilst eq. (3) ensures thermal equilibrium at inverse temperature β, it does not fully determine the form of K. Typically $K(\mu \to \nu)$ is the combination of an attempt frequency to move to state ν given that the system is in state μ, multiplied by an acceptance probability. For simplicity of exposition we will take all the attempt frequencies to be equal to unity, so that K is just an acceptance probability and so must obey $0 \le K \le 1$. The Metropolis algorithm[7] is fully specified by requiring that $K(\Delta E) = 1$ for $\Delta E < C$, with C as large as possible (maximising acceptance rates) leading to

$$K_{\text{Metropolis}}(\Delta E) = \min\left(1, e^{-\beta \Delta E}\right), \tag{4}$$

whereas the Glauber acceptance function [8] arises by requiring that $K(\Delta E) + K(-\Delta E) = 1$, leading to

$$K_{\text{Glauber}}(\Delta E) = 1/\left(1 + e^{\beta \Delta E}\right). \tag{5}$$

These are compared graphically in Fig. 1.

We now consider the case when the true energy change is not known exactly, and we must accept moves with probabilty $A(x)$ where x is only an *estimate* of the energy change. We will assume that these estimates have statistically independent errors. If $f(x|\Delta E)$ is the probability density of estimating that the energy change is x when its true value is ΔE, then the net probability of accepting a move whose true energy change is ΔE is given by

$$K(\Delta E) = \int_{-\infty}^{\infty} f(x \mid \Delta E)\, A(x)\, dx. \tag{6}$$

It is our aim to choose $A(x)$ such that $K(\Delta E)$ satisfies detailed balance (3).

We can 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 aproximation to eq. (3) of great value in optimisation. The resulting acceptance function K when f is given by a Gaussian distribution is simply related to the standard Error Function and is graphed in Fig. 1. The resemblance between this and the Glauber

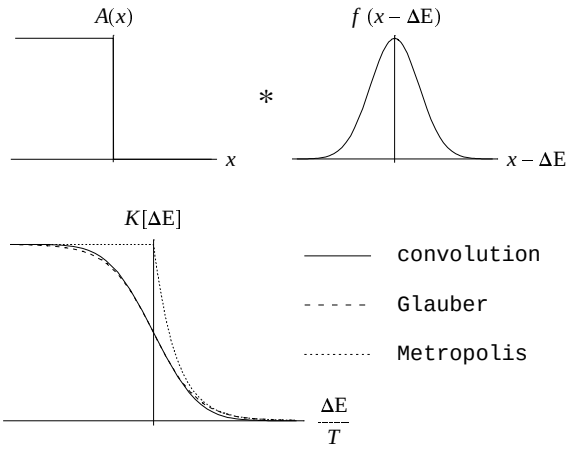

FIG. 1: A simple approximate stochastic annealing is obtained by accepting moves on the basis of the sign of the estimated energy change x. The acceptance probability as a function of the true underlying energy change ΔE is then given by a convolution with the error distribution. As shown for the Gaussian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

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, so detailed balance is not strictly achieved. The standard deviation σ of the Gaussian distribution controls an approximate effective temperature using this rule, as inverting eq. (3) gives

$$T \simeq \sqrt{\frac{\pi}{8}}\sigma\left(1 - 0.018\,(\Delta E/T)^2 + \text{order}\,(\Delta E/T)^4\right). \quad (7)$$

For many optimisation purposes the departure from detailed balance, due to the energy dependent terms at large $|\Delta E/T|$, is not a problem. Fig. 2 shows how we successfully used this approach to optimise 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 7. In another paper[9] we show that this approach can be exploited to unravel a benchmark problem in stochastic optimisation, the probabilistic travelling salesman problem [10, 11].

We now attempt directly to choose $A(x)$ properly, such that K exactly satisfies detailed balance, eq. 3, assuming that the error distribution $f(x|\Delta E)$ is fully known. We first note that this leads to some bounds on the behaviour of the right tail of $A(x)$ and a left tail of $f(x|\Delta E)$.

Given that $K(\Delta E) \leq 1$ even for negative arguments, the detailed balance condition requires $\int_{-\infty}^{\infty} f(x \mid \Delta E)\, A(x)\, dx \leq e^{-\beta \Delta E}$ and for large positive ΔE this severely restricts all contributions to the LHS. First consider $x \simeq \Delta E$ for which $f(x \mid \Delta E)$ is not expected to be small. Then the exponential decay on this contribution must come from $A(x)$ falling off at least as fast as $e^{-\beta x}$ as $x \to \infty$, which becomes important in our later analysis. Second consider $x \lesssim 0$, for which $A(x)$ cannot become small or we would be heavily rejecting even moves which appear to be downwards in energy. Then the exponential decay on this contribution must come from $f(x|\Delta E)$, $x \lesssim 0$, falling off at least as fast as $e^{-\beta \Delta E}$ for $\Delta E \to \infty$. This latter sets a fundamental restriction on our stochastic annealing: the probability for large energy increases to be estimated as negative must fall off faster than a

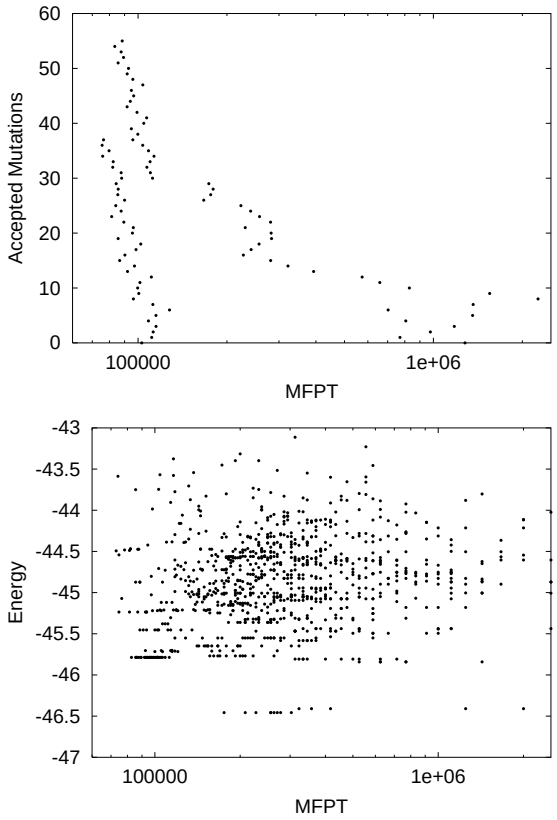

FIG. 2: Stochastic annealing applied to protein folding, in a simple cubic lattice model. The chains were 27 units long and the space of amino acid sequences was explored for the ability of the molecule to fold spontaneously to a fixed $3 \times 3 \times 3$ cubic target conformation. (a) The upper panel 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 analogue 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 indivi-diual runs of Stochastic Annealing find folding speeds approaching the fastest available, and much better than would be achieved by seeking minimum energy (the Shaknovitich scheme [12]).

thermal Boltzmann factor.

Now we begin direct analysis of the detailed balance condition (3), inroducing the substitutions $f(x|\Delta E) = e^{\frac{\beta}{2}(x-\Delta E)}g(x,\Delta E)$ and $A(x) = e^{-\frac{\beta x}{2}}a(x)$ so that eqs. (3) and (6) combined simplify down to

$$\int_{-\infty}^{\infty} a(x)\ (g(x,\Delta E) - g(x,-\Delta E))\ dx = 0. \tag{8}$$

Without significant loss of generality we can try

$$a(x) = \int_{-\infty}^{\infty} h(\Delta E')\, j(x,\Delta E')\, d\Delta E' \tag{9}$$

where h and j are functions to be chosen. Substituting this into equation 8 we find

$$\int_{-\infty}^{\infty} h(\Delta E')\left[k(\Delta E,\Delta E') - k(-\Delta E,\Delta E')\right]\, d\Delta E' = 0 \tag{10}$$

where

$$k(\Delta E,\Delta E') = \int_{-\infty}^{\infty} g(x,\Delta E) j(x,\Delta E')\, dx. \tag{11}$$

Then eq. 10 is satisfied when $h(\Delta E') = h(-\Delta E')$ and $k(\Delta E,\Delta E') = k(-\Delta E,-\Delta E')$ are even functions in the given sense. We have not succeeded in taking this most general case significantly further, the hard part being to implement $0 \le A(x) \le 1$ for a probability.

For further progress we now specialise to the case of invariant error distributions, where $f(x|\Delta E) = f(x-\Delta E)$, meaning that the distribution of error is independent of ΔE itself. For this case, we can choose $j(x,\Delta E') = g(x-\Delta E')$ and then from eq. 11 we find $k(\Delta E,\Delta E')$ is a suitably even function. Then we have a solution to eq 8 given by

$$a(x) = \int_{-\infty}^{\infty} h(\Delta E')\, g(x-\Delta E')\, d\Delta E' \tag{12}$$

provided $h(x)$, and correspondingly $\widetilde{h}(p)$ below, is even.

We now aim to choose h to maximise the move acceptance rates, which are governed by $A(x)$. We follow the Metropolis methodology in choosing $A(x)$ $(= e^{-\frac{\beta x}{2}}a(x))$ to be identically 1 below some threshold, $x < C$. Then using the Fourier-Laplace transform defined by $\widetilde{f}(p) = \int_{-\infty}^{\infty} e^{-px} f(x)\, dx$, the transform of a becomes

$$\widetilde{a}(p) = \int_{-\infty}^{C} e^{\frac{\beta x}{2}} e^{-px}\, dx + \int_{C}^{\infty} b(x) e^{-px}\, dx \tag{13}$$

where $b(x) = A(x)e^{\frac{\beta x}{2}}$ for $x \ge C$, $b(x) = 0$ for $x < C$. From eq.(12) the transform of the new function $b(x)$ is given by

$$\widetilde{b}(p) = \widetilde{h}(p)\widetilde{g}(p) - \frac{e^{\left(\frac{\beta}{2}-p\right)C}}{\frac{\beta}{2}-p}. \tag{14}$$

The exponential bound we established on the right tail of $A(x)$ implies that $\widetilde{b}(p)$ is bounded for all $\mathrm{Re}\, p > -\frac{\beta}{2}$, whereas the integrability of $f(x)$ together with the exponential bound on its left tail only imply that $\widetilde{g}(p)$ is bounded for $-\frac{\beta}{2} \le \mathrm{Re}\, p < \frac{\beta}{2}$. Thus $\widetilde{h}(p)$ must be chosen to cancel both any divergences of $\widetilde{g}(p)$ in $\mathrm{Re}\, p \ge \frac{\beta}{2}$ and the apparent pole at $p = \beta/2$. We first turn to the Wiener-Hopf method [13] to define

$$\widetilde{g}(p) = \widetilde{g_L}(p)\widetilde{g_R}(p) \tag{15}$$

where $\widetilde{g_L}(p)$ is bounded and non-zero for $\mathrm{Re}\, p < \frac{\beta}{2}$, and $\widetilde{g_R}(p)$ is bounded and non-zero for $\mathrm{Re}\, p > \frac{\beta}{2}$; it also follows from the bounded window for $\widetilde{g}(p)$ that $\widetilde{g_R}(p)$ is bounded for the wider range $\mathrm{Re}\, p > -\frac{\beta}{2}$. Then by choosing $\widetilde{h}(p) = \frac{B}{\left((\beta/2)^2-p^2\right)}\frac{1}{\widetilde{g_L}(p)\widetilde{g_L}(-p)}$, we can ensure that $\widetilde{h}(p)\widetilde{g}(p) = \frac{B}{\left((\beta/2)^2-p^2\right)}\frac{\widetilde{g_R}(p)}{\widetilde{g_L}(-p)}$ is duly bounded for $\mathrm{Re}\, p > -\frac{\beta}{2}$ except for the (desired) pole at $p = \frac{\beta}{2}$. Choosing the constant $B = \beta\frac{\widetilde{g_L}\left(-\frac{\beta}{2}\right)}{\widetilde{g_R}\left(\frac{\beta}{2}\right)}$ makes the residue of this pole cancel when we reassemble the expression (14) for $\widetilde{b}(p)$ to give

$$\widetilde{b}(p) = -\frac{e^{\left(\frac{\beta}{2}-p\right)C}}{\frac{\beta}{2}-p} + \frac{\beta}{\left((\beta/2)^2-p^2\right)}\frac{\widetilde{g_R}(p)\widetilde{g_L}\left(-\frac{\beta}{2}\right)}{\widetilde{g_R}\left(\frac{\beta}{2}\right)\widetilde{g_L}(-p)}. \tag{16}$$

This should be the optimal solution. We cannot incorporate any new factors into $\widetilde{h}(p)$ because, being even, they would have to be bounded both for $Re(p) > -\frac{\beta}{2}$ and for $Re(p) > \frac{\beta}{2}$, and we would run up against the limitations of Liouville's Theorem. The parameter C might appear to be a remaining degree of freedom, but it drops out of the acceptance function itself which is given directly in terms of

$$\widetilde{a}(p) = \widetilde{h}(p)\widetilde{g}(p) = \frac{\beta}{\left((\beta/2)^2-p^2\right)}\frac{\widetilde{g_R}(p)\widetilde{g_L}\left(-\frac{\beta}{2}\right)}{\widetilde{g_R}\left(\frac{\beta}{2}\right)\widetilde{g_L}(-p)}. \tag{17}$$

The parameter C drops out because it simply reflects the partition we introduced in equation (13). The feature which we have not strictly guaranteed is that $A(x) = 1$ for $x < C$ rather than some different cutoff.

As an example of our approach above, consider the simple case where

$$f(x-\Delta E) = \frac{\gamma}{2}e^{-\gamma|x-\Delta E|} \tag{18}$$

which leads to $\widetilde{g}(p;\gamma) = \frac{\gamma^2}{(\gamma-\frac{\beta}{2}-p)(\gamma+\frac{\beta}{2}+p)}$. The choice of $\widetilde{g_L}$ and $\widetilde{g_R}$ is trivial by inspection, giving $\widetilde{a}(p) = \frac{\beta}{(\beta/2)^2-p^2}\frac{\gamma+\beta}{\gamma}\frac{\gamma-\frac{\beta}{2}+p}{\gamma+\frac{\beta}{2}+p}$. The inverse transformation can also be performed by inspection to give

$$A(x) = \min\left(1, \frac{\gamma^2-\beta^2}{\gamma^2}e^{-\beta x} + \frac{\beta^2}{\gamma^2}e^{-(\gamma+\beta)x}\right). \tag{19}$$

We note that this acceptance function only remains positive for $x \gg 0$ when $\gamma \geq \beta$, which is precisely the limit of achievable stochastic annealing discussed earlier, and the case $\gamma \to \infty$ also duly recovers the straightforward Metropolis method. We have analysed 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 $\widetilde{g}(p)$ has no obvious Wiener-Hopf factorisation. 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 taking the solvable form of multiply convolved exponentials and (by the Central Limit Theorem) approaching Gaussian form as $N \to \infty$. In this case $\widetilde{g}(p) = \widetilde{g}(p;\gamma)^N$ with $\gamma = \sqrt{2N}/\sigma$, where σ is the standard deviation of x, leading to $\widetilde{a}(p) = \frac{\beta}{(\beta/2)^2-p^2}\left(\frac{\gamma+\beta}{\gamma}\frac{\gamma-\frac{\beta}{2}+p}{\gamma+\frac{\beta}{2}+p}\right)^N \to \frac{\beta}{(\beta/2)^2-p^2}e^{\beta(p-\beta/2)\sigma^2/2}$ as $N \to \infty$ at fixed σ. The corresponding acceptance function by inverse transformation is then given by

$$A(x) = \min\left(1, e^{-\beta(x+\beta\sigma^2/2)}\right) \qquad (20)$$

The optimal acceptance rules found above all obey the bounds that $0 \leq A(x) \leq 1$ required for a probability, but they only do so *a postiori* so it can be objected that our optimal method gives no guarrantee of this outcome *a priori*. We have discovered a general but sub-optimal solution for the acceptance function which does assure the requirement. The key to this solution is to note that, given the result 17, our requirements on the factorisation of $\widetilde{g}(p)$ can be relaxed to require only that $\widetilde{g_R}(p)$ is bounded for $\mathrm{Re}\, p > -\frac{\beta}{2}$ and that $\widetilde{g_L}(p)$ is non-zero for $\mathrm{Re}\, p < \frac{\beta}{2}$. Then assuming that $f(x) = 0$ for $x < C$, an acceptable factorisation is given by $\widetilde{g_R}(p) = \widetilde{g}(p)e^{pC}$, $\widetilde{g_L}(p) = e^{-pC}$ leading directly to $\widetilde{a}(p) = \widetilde{h}(p)\widetilde{g}(p) = \frac{\beta}{((\beta/2)^2-p^2)}\frac{\widetilde{g}(p)}{\widetilde{g}(\frac{\beta}{2})}$, at which point we can let $C \to -\infty$ so no significant new restriction has been imposed on f. The resulting acceptance function by inverse transformation can be expressed as the convolution

$$A(x) = \frac{1}{\widetilde{f}(\beta)}\int_{-\infty}^{\infty} K_{\mathrm{Metropolis}}(x-y)\; e^{-\beta y} f(y)\; dy. \quad (21)$$

It can be verified by direct substitution that this does obey the detailed balance condition (3), it is manifestly positive definite, and it gives maximum acceptance 1 as $x \to -\infty$. Simple comparison shows the resulting acceptance probabilities are lower than the optimal values which we have calculated these, and the key difference is that an interval where $A = 1$ is no longer imposed.

Given the wide impact of equilibrium statistical mechanics, we are confident that our new method of exactly thermal stochastic annealing will find significant direct application. The chief limitation is that the error distribution must be known. In this context the Gaussian case is particulary important as it can be approached simply through multiple sampling, although its variance is also required. Estimating the variance from the same sample does induce errors, but these can be quantified and become negligible as the sample size becomes large[14]. It is a more open question whether it is viable to estimate by measurement the full distribution of error and apply this to eq. (21).

Underpinned by the exact results our analysis provides a powerful new tool in stochastic optimisation. 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! In the protein folding problem we used limited samples of folding time for these estimates, making the high temperature part of the annealing computationally the cheapest. This interestingly complicates the long noted problem of how to choose the optimum cooling schedule, because cooling by reducing sample errors introduces computational costs growing as $1/T^2$ per move.

NEB would like to thank BP and EPSRC for the support of a CASE award during this research.

* Electronic address: r.c.ball@warwick.ac.uk; URL: `http://www.phys.warwick.ac.uk/theory/`
† Electronic address: tmf20@cam.ac.uk; URL: `http://www.tcm.phy.cam.ac.uk/~tmf20/`
‡ Electronic address: Neill.Bowler@physics.org; URL: `http://uk.geocities.com/neill_bowler`; Present address: Met Office, Maclean Building, Crowmarsh-Gifford, Oxfordshire OX10 8BB, UK.

[1] S. Kirkpatrick, C. D. Gelatt Jr and M. P. Vecchi, *Science* **220**, 671 (1983).
[2] A. Harju *et al.*, *Phys. Rev. Lett.* **79**, 1173 (1997).
[3] Edward J. Anderson and Michael C. Ferris, *SIAM J. Optimiz.* **11**, 837 (2001).
[4] H. Robbins and S. Munro, The annals of mathematical statistics **22**, 400 (1951).
[5] W. B. Gong, Y. C. Ho, and W. Zhai, in *Proceedings of the 31st IEEE conference on decision and control* (IEEE, PO Box 1331, Piscataway, NJ, 1992), pp. 795–802.
[6] A. A. Bulgak and J. L. Sanders, in *Proceedings of the 1988 Winter Simulation Conference* (IEEE, PO Box 1331, Piscataway, NJ, 1988), pp. 684–690.
[7] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[8] R. J. Glauber, J. Chem Phys **4**, 294 (1963).
[9] N. E. Bowler, T. M. Fink, and R. C. Ball, Characterization of the probabilistic travelling salesman problem, submitted to Physical Review E.
[10] P. Jaillet, *Operations Research* **V36**, 929, (1988).

[11] P. Jaillet, *Math. Oper. Res.* **18**, 51 (1993).
[12] A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **92**, 1282 (1995).
[13] P. M. Morse and H. Feshbach, in *Methods of theoretical physics* (McGraw-Hill, London, 1953), pp. 978–980.
[14] N.E. Bowler, PhD Thesis, University of Warwick, UK (2001).