Pramāṇa – J. Phys., Vol. 36, No. 1, January 1991, pp. 1–77. © Printed in India.

Spin glass, the travelling salesman problem, neural networks and all that

G VENKATARAMAN and G ATHITHAN
ANURAG, Kanchanbagh, Hyderabad 500 258, India

MS received 19 October 1990

Abstract. This paper presents an overview of diverse topics that are seemingly different but interrelated, with strong connections to statistical mechanics on the one hand and spin glass physics on the other. Written primarily for an inter-disciplinary audience, we start with a brief recapitulation of the relevant aspects of statistical mechanics, particularly those needed for understanding the recently-popular simulated-annealing technique used in optimization studies. Then follows a survey of the spin glass problem, with particular attention to the consequences of quenched randomness. The travelling-salesman problem is considered next, as also the impact made on it by the spin glass problem. Several examples are then presented of optimization studies wherein the simulated-annealing concept has been profitably used. Attention is also drawn in this context to the lessons provided by the spin glass problem. Finally, a brief survey of neural networks is made, essentially from a physicist's point of view. The different learning schemes proposed are discussed, and the relevance of spin models and their statistical mechanics is also discussed.

Keywords. Spin glass; travelling salesman problem; neural networks; simulated annealing.

PACS Nos 05·50; 64·60

Table of contents

1. Introduction

The title of this article is intended to highlight interesting correspondences drawn in recent times between diverse, seemingly—disconnected fields. An important outcome has been that physicists have begun to venture into areas not traditionally their preserve; for their part, non-physicists appear to be drawing useful lessons from condensed matter physics. If there is one common theme underlying all these developments, it is optimization.

In condensed matter physics the optimization is with respect to free energy, and one now has a reasonably good idea of the different states of matter possible under various conditions. Thus one is familiar with the liquid state, the glassy state prepared by a rapid quench, and the crystalline state resulting from careful and controlled slow-cooling. One further knows that while glass represents a metastable state, a crystal, by contrast, represents a state in equilibrium. A new-comer to this family of phases is the spin glass on which the spotlight will be turned in a later section.

The requirement of optimization occurs not only in condensed matter physics but in a variety of other fields like engineering, biology and so on. Frequently, the problems feature many degrees of freedom (as in condensed matter physics, though not in the same range of numbers) as well as a conflict of interest, analogous to the *frustration* one is familiar with in random systems. Thus it is that spin glass has begun to attract attention among non-physicists.

The plan of the article is as follows: We start with a few basics of statistical mechanics, together with a brief discussion of the celebrated Metropolis algorithm (Metropolis *et al* 1953) which facilitates numerical simulation of statistical mechanical systems. Lately, this algorithm has begun to attract wide attention in the context of optimization problems. Section 3 is devoted to a comprehensive survey of the spin glass problem and the message it holds. Follows then in § 4 a survey of the Travelling Salesman Problem (TSP) and the application of the Metropolis algorithm to it. The statistical mechanical aspects of the TSP are also surveyed, as well as their relationship to ›pin glass. The use of the Metropolis algorithm in various other optimization problems is briefly reviewed in § 5. Nature is perhaps the biggest practitioner of optimization; a brief contact is therefore made in § 6 with optimization studies in the biological field. In Nature's scheme of things, the brain occupies a unique place. The manner in which it works is fascinating, and numerous have been the efforts to unravel its mysteries. Engineers try to mimic the brain with electrical circuits referred to as *neural networks.* Recently, statistical mechanics has been applied to neural networks, a development which is surveyed in § 7. The paper concludes with general remarks arising out of all the recent experience.

2. Statistical mechanics and the Metropolis algorithm

Statistical mechanics is a subject with much intellectual appeal, having its origins in efforts to provide a microscopic foundation for thermodynamics. It will be recalled that in the latter, one is interested in the equilibrium state of a macroscopic system in the presence of various external forces like hydrostatic pressure, magnetic field etc. The equilibrium state is obtained by minimising an appropriate free energy (the Helmholtz free energy, the Gibbs free energy etc.). Indeed, chemists and metallurgists make extensive use of the free energy concept in discussing phase diagrams.

Statistical mechanics teaches us how the thermodynamic free energy F is related to microscopic physics. Here we will be concerned mainly with the Helmholtz free energy $F = E - TS$. At the microscopic level one has a Hamiltonian $\mathscr{H}$ with attendant equations of motion, and F is related to $\mathscr{H}$ by

$$F = -k_B T \operatorname{In} Z \tag{1}$$

where Z the partition function is given by

$$Z = \operatorname{Tr} \exp(-\beta\mathscr{H}), \quad \beta = 1/k_B T. \tag{2}$$

In (2), Tr implies a sum over all the allowed states of the system. We shall be concerned only with classical systems. If X comprehensively denotes the configuration coordinates and $X_1, X_2 \cdots X_N$ those appropriate to the allowed states $1, 2, \cdots N$, then

$$Z = \sum_{i=1}^{N} \exp\{-\beta\mathscr{H}(X_i)\}. \tag{3}$$

For discussing phase changes, one usually follows Landau and adopts a phenomenological approach. To recall, one supposes that the free energy is given by an expansion of the form

$$F = F_0 + a\psi^2 + b\psi^4, \tag{4}$$

where ψ is the order parameter,

$$b > 1 \text{ and } a = \alpha(T - T_c), \quad \alpha > 1. \tag{5}$$

The curves corresponding to (4) are shown in figure 1 from which it is clear that for

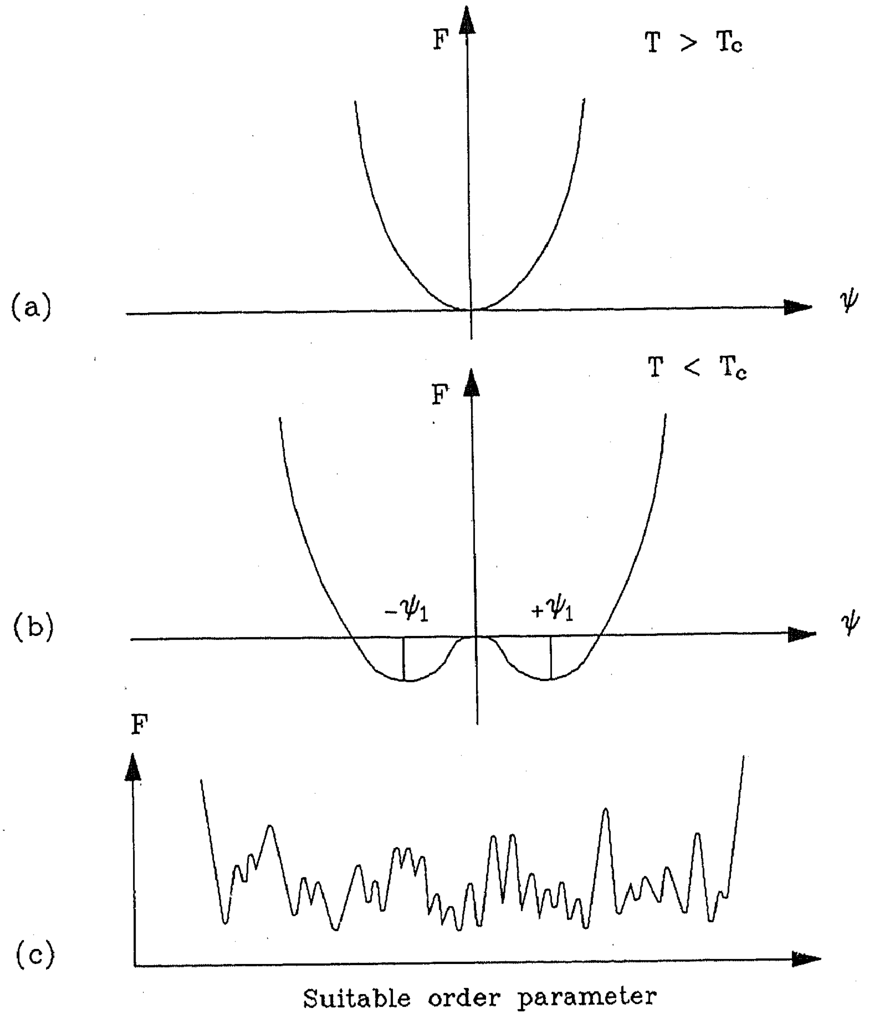


Figure 1. Curves (a) and (b) depict the familiar Landau free energy curves (see eq. (4) and the discussion following that). The equilibrium state of the system is that corresponding to the minimum of the free energy. A similar plot for spin glass is shown in (c). The abscissa for case (c) is unspecified, but denotes an appropriate coordinate. As we shall see in § 3.8, there is really an infinity of such variables! Observe the rough terrain. Its nature will receive attention later.

$T > T_c$ the order parameter $\psi = 0$ while for $T < T_c$ there are two options $\psi = \pm\psi_1$. The latter states are ordered since the order parameter is non-vanishing.

As in the case of fundamental symmetries in particle physics, one does not know *a priori* what the relevant order parameter is for a given phase transition. The problem is especially acute if the phase transition is of a type not known earlier (as happened in the case of spin glass). Knowledge of the candidate order parameter set usually emerges from experimental facts, supplemented by inspired guessing!

The situation for spin glass corresponding to figure 1(b) is shown in 1(c), and we reserve for later narration the fascinating story about how such a picture emerged. For the moment let us go back to traditional equilibrium statistical mechanics and to standard problems where no phase change is involved (like liquid-state properties, for example). Here statistical mechanics no doubt provides a complete set of formal rules, but at the practical level things are a bit disappointing since the calculation involves the evaluation of horrendous multi-dimensional integrals. For instance, if we consider a classical gas, the integrals to be performed are of the form

$$\int d^{3N}p\,d^{3N}q(\cdots)$$

where $(d^{3N}p\,d^{3N}q)$ is an element of the 6 N-dimensional phase space of the system. Except for highly idealized models, such integration is impossible to perform even in supercomputers since $N \sim 10^{23}$.

A classic paper by Metropolis *et al* (1953) showed a way out of this impasse, and understandably has become a bench mark. There are two important messages in this paper. First is that one could (at least under certain circumstances) obtain physically meaningful results by considering as few as, say, a thousand particles, instead of an Avagadro number of them. Secondly, the integration difficulty can be further simplified by integrating over a random sample of points in phase space rather than over all of it.

At the practical level, the Metropolis algorithm involves the steps outlined below. Let X denote collectively the configuration coordinates of a system at temperature T and $E(X)$ the energy associated with it.

Step 0: Change X to $X+\Delta X$ where ΔX is a small, arbitrary displacement.
Step 1: Calculate $E(X+\Delta X)$.
Step 2: Compute the difference $\Delta E = E(X+\Delta X) - E(X)$.
Step 3: If $\Delta E < 0$, i.e., if the move $X \rightarrow X+\Delta X$ lowers the energy, accept the new configuration and go back to Step 0; if not, proceed to Step 4.
Step 4: Select a random number R between 0 and 1 and compare it with $P = \exp(-\Delta E/k_B T)$. If $R \leqslant P$, accept the new configuration $X+\Delta X$ and go back to Step 0. If $R \leqslant P$, retain the configuration X and start all over again from Step 0.

Steps 0 to 3 do not require explantion. Step 4 has the significance that even if the change X to $X+\Delta X$ costs an *increase* in energy, it has a certain probability of occurring. In particular, the acceptance scheme prescribed via the random number R ensure detailed balance.

For obvious reasons, the above algorithm is also frequently referred to as the *Monte Carlo method.* In its practical implementation, one must allow for a sufficient number of Monte Carlo steps for the system to settle down to equilibrium (at temperature T) before extracting results. Thereafter, the point representing the state of the system wanders in phase space in a manner required by equilibrium statistical mechanics. All the desired results may be obtained by suitably averaging over this trajectory. A schematic illustration of this process is given in figure 2. Further elucidation is available in figure 3 which amplifies the famous ergodic hypothesis of Gibbs.

It is necessary to be clear about the difference between the role played by the free energy F and the (internal) energy states E_1, E_2...etc. of the Hamiltonian. At any given temperature T, the system always tries to minimize F, which is not the same as finding the lowest energy state in the set $\{E_1, E_2 \ldots\}$. In fact, when $T \neq 0$, the system continuously wanders amongst all the possible energy states, although it spends more time amongst the lower energy ones (see figure 4). However, at $T=0$

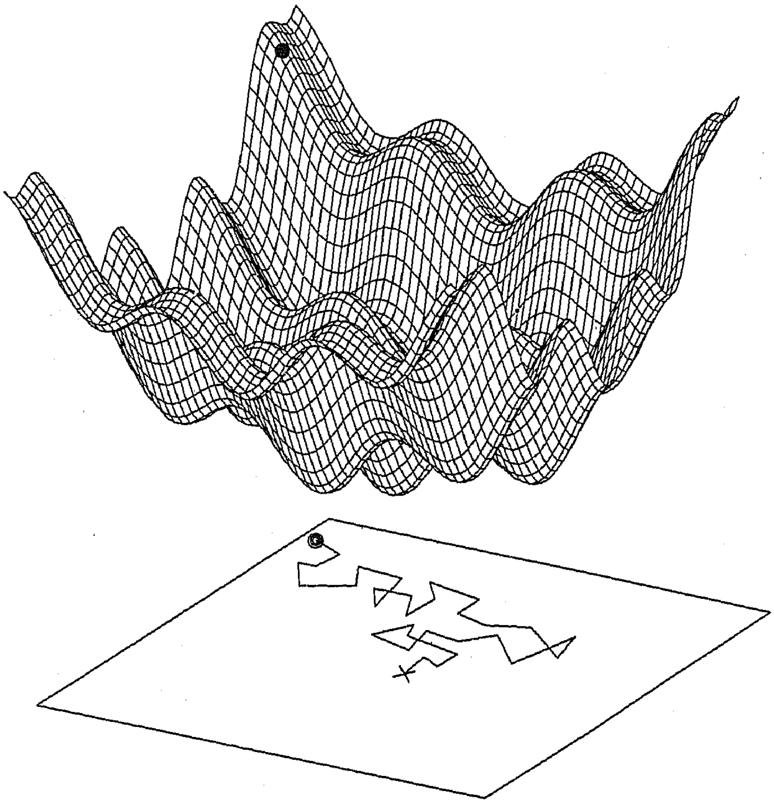


Figure 2. Schematic plot of a two-dimensional energy landscape $E(X_1, X_2)$. There are hills and valleys, and the system wanders over this terrain which projects as a random walk in (X_1, X_2) space—see the lower half of the figure.

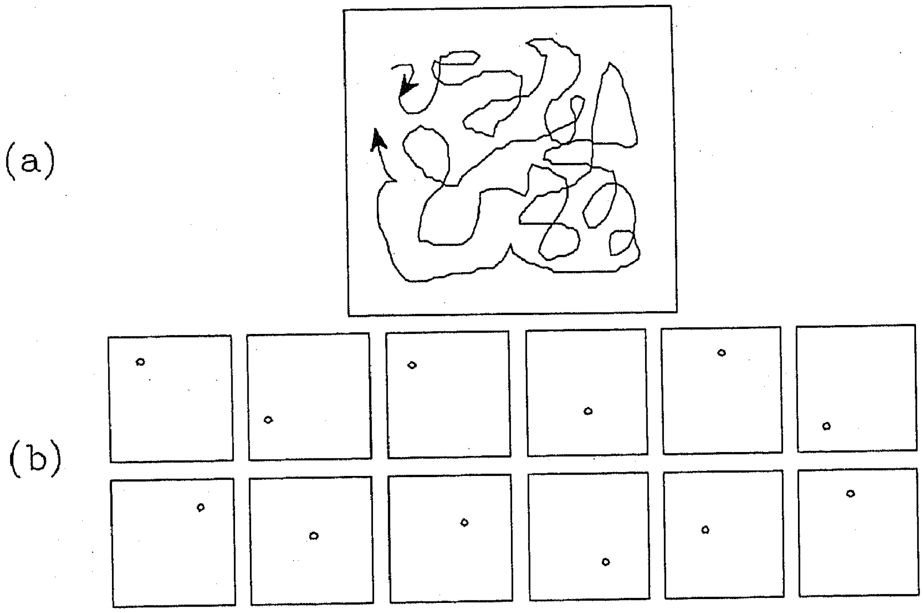


Figure 3. Illustration of the ergodic hypothesis. The time evolution of the system can be depicted as a trajectory as in (a); recall also the previous figure. If one waits long enough, the trajectory would cover the whole of phase space. Physical properties may be evaluated as time averages over this trajectory. In (b), we have several copies of the system, all observed at the same instant of time. The points represent the state of the system in the various members of the ensemble at that instant. The ergodic hypothesis states that for a system in thermodynamic equilibrium, time average equals the ensemble average. Subtler aspects of the ergodic hypothesis are discussed by Ma (1975).

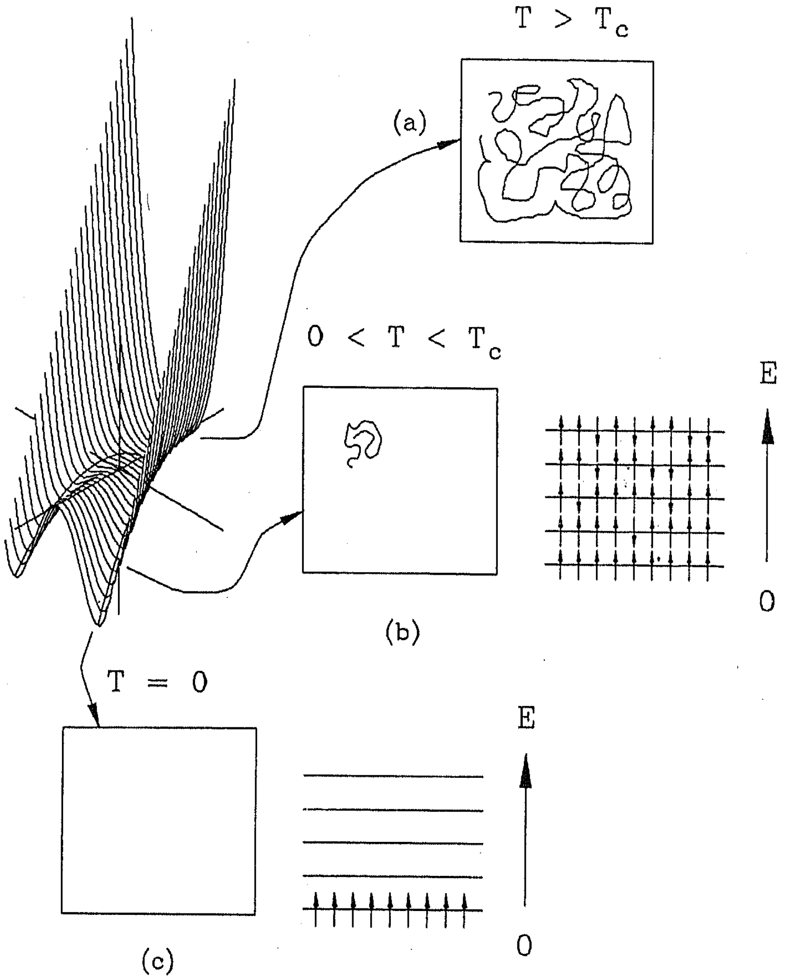


Figure 4. Illustration of the difference in the roles played by the free energy and the internal energy. The sketches are for a ferromagnetic Ising spin system. At each temperature, the free energy is always minimized but that is the combined effect of different spin states with different (internal) energies. The phase space (really a N-dimensional one!) trajectory depicts the exploration of all these various possible spin configurations. At 0°K, the system settles down in the state of lowest internal energy, corresponding to perfect spin alignment. There is then a one-to-one correspondence between E_{min} and F_{min}.

the system settles down permanently at the state of lowest energy E_{min}, and only then does the distinction between F_{min} and E_{min} vanish.

The difference between F and E is stressed because in optimization problems outside the domain of physics, one does not really have a F (although, lately, statistical mechanics is being injected into such problems, as we shall see in §4). The optimization in such problems often boils down to finding the minimum energy state which, in thermodynamic language, corresponds to a 0°K situation.

One must also put in perspective certain facts related to the technique of *quenching* and *annealing* frequently resorted to by materials scientists (see figure 5). Quenching involves step cooling, the sudden change in temperature causing the system to rapidly coast down the free energy landscape, almost invariably getting trapped in a local

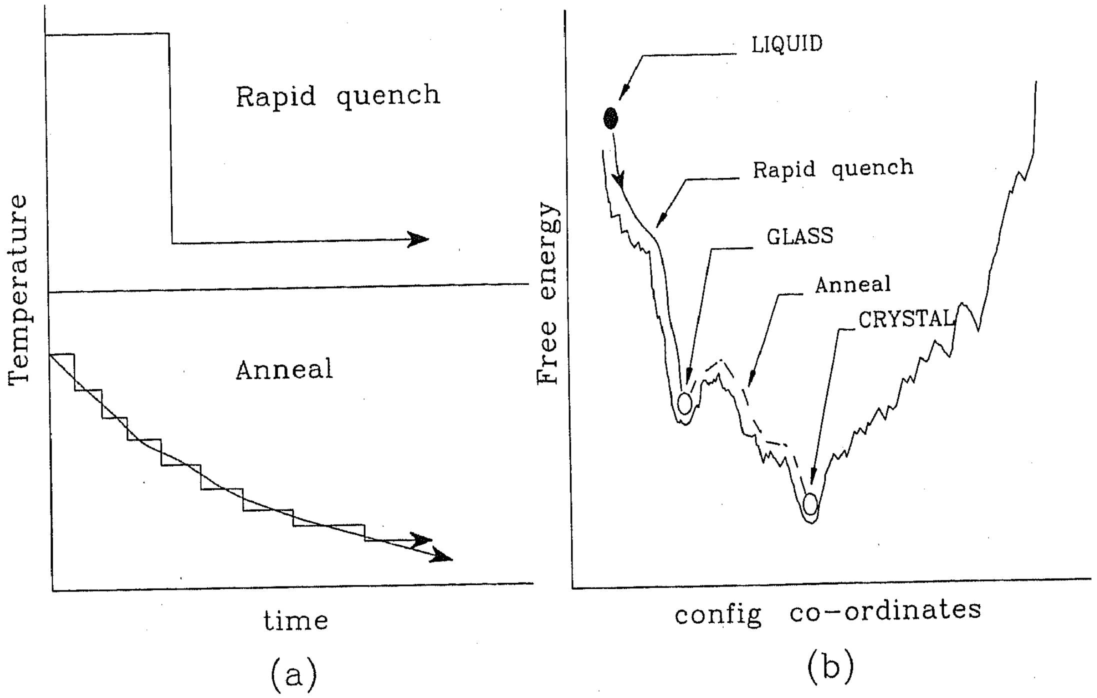


Figure 5. In (a) is shown the time-temperature diagram for quench as well as for anneal. In the latter case, the temperature decrease can be continuous or step-like. The familiar example of glass formation by rapid quench and devitrification by annealing is illustrated in (b).

valley during the downhill slide. Sometimes quenching endows the material so treated with beneficial properties, which is why the blacksmith resorts to it. On the other hand, quenching also leads to defects (if not, at times, total disorder) which may not be always desirable. In such case one resorts to annealing, a process which provides an escape route from local valleys, should trapping occur. In a nutshell, annealing can undo what quenching does. For examples, metallic glasses are prepared by rapid quenching (Cahn 1980). However, prolonged annealing can convert them into the crystalline form, a process known as devitrification. This is an example of annealing assisting the system out of a metastable state (local valley) and enabling it to find a stable state (global minimum). As we shall see later, this idea has been successfully applied in many optimization problems (through the use of the Metropolis algorithm). Interestingly, such problems often have many features in common with spin glass so that, spurred by developments in the spin glass field, many optimization problems are being now studied both from the statistical mechanical point of view as well as the numerical, the latter studies often being for corroborating the results obtained in the former. It is the objective of this article to highlight some of these recent developments, an exercise we start by considering first the spin glass problem.

3. The spin glass problem

3.1 *Introduction*

In this section we survey the spin glass problem, keeping in mind the other topics to

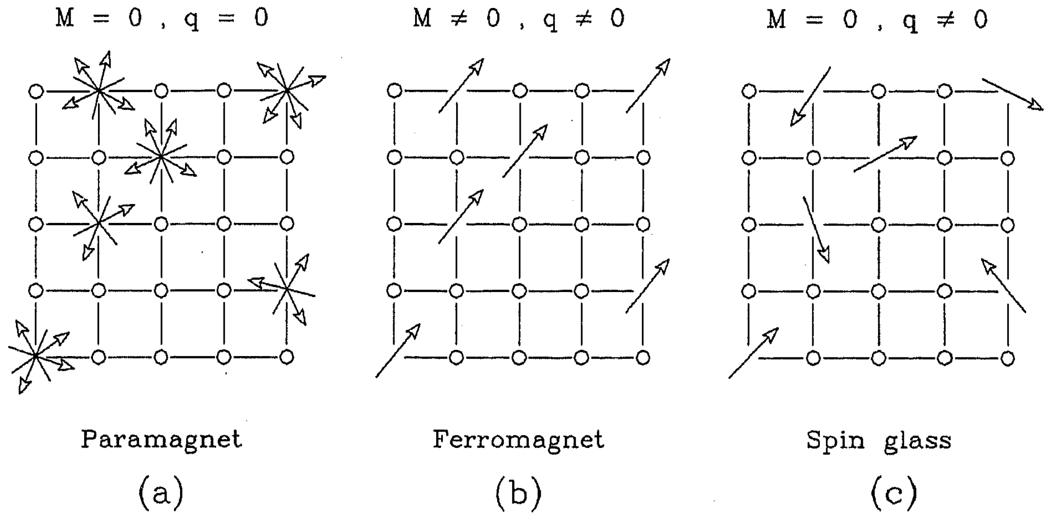


Figure 6. Differences between a paramagnet, ferromagnet and spin glass. Here $M = (1/N)\Sigma_i m_i$ and $q = (1/N)\Sigma_i m_i^2$, where m_i is the magnetization at site i. Observe that $q \neq 0$ for spin glass.

be discussed later. An early review on this subject is by Ford (1982) while subsequent developments have been summarized by many, for example, by Southern (1987) and by Williams (1987). An in-depth analysis can be found in the paper of Binder and Young (1986).

As the name itself implies, the term spin glass is applied to a material that exhibits a magnetic characteristic somewhat analogous to that displayed by its more familiar counterpart namely, glass. The distinguishing feature of the latter is the presence of atomic disorder. No doubt the same is also true of a liquid but whereas the atoms are mobile in a liquid, in a glass their positions are frozen. There is a corresponding difference between the spins in a paramagnet and those in a spin glass as schematically illustrated in figure 6.

Spin glass behaviour is observed in a wide variety of materials, of which AuFe and CuMn alloys are regarded as archetypal. In substances of the above type, one has typically 1–10 atomic % of a magnetic impurity (e.g. Fe) in a non-magnetic host (e.g. Au). As the substances is cooled, the (magnetic) susceptibility exhibits a peak (or rather a cusp) similar to that seen for various properties during an order-disorder transformation. It is now generally accepted that the above mentioned peak in the susceptibility signals the onset of a new phase at lower temperatures, the spin glass phase.

The basic reasons for the occurrence of such a phase are also understood, and may be appreciated by referring to figure 7. From figure 7(a) we see that since the magnetic atoms enter the host lattice at random sites, the distances between the spins also vary in a similar manner. Next, the interaction between the magnetic atoms is of the famous RKKY (Ruderman–Kittel–Kasuya–Yosida) form, having an oscillatory nature as in figure 7(b). Together, figures 7(a) and (b) imply that not only does the magnitude of the nearest-neighbour exchange interaction vary randomly, but also that its *sign fluctuates*. The latter is significant, since a positive exchange interaction favours the parallel alignment of spins while a negative interaction favours the opposite. Thanks to the inherent randomness, it becomes difficult in these materials to find an

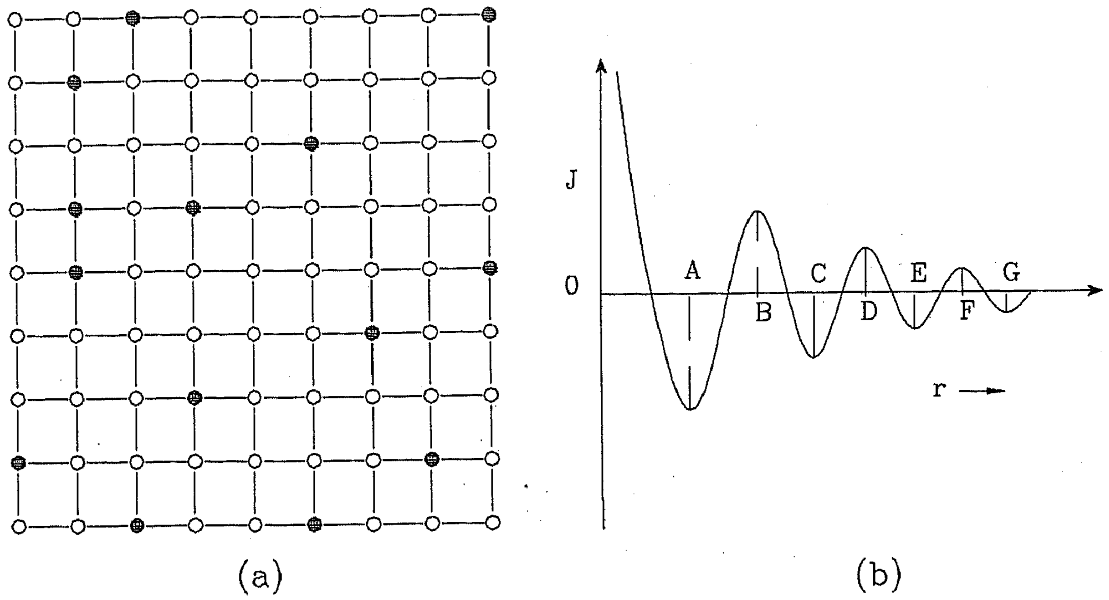


Figure 7. (a) Schematic representation of a spin glass alloy. The host atoms are shown as open circles and the magnetic impurities as solid circles. (b) RKKY interaction $J(r)$ between two magnetic atoms as a function of the separation r between them. At $r = \mathrm{OA}$, OC, OE and OG, the coupling is antiferromagnetic while for $r = \mathrm{OB}$, OD and OF, it is ferromagnetic.

overall spin alignment that can *simultaneously* satisfy *all* the exchange bonds. This is the famous *frustration* problem, first highlighted by Anderson. (In his article, Ford recalls the history of this terminology.) Thus the two key elements required for spin glass behaviour are (i) randomness and (ii) frustration. In general, "frustration means competing interactions and therefore unsatisfactory states, degeneracy, metastability, and sensitivity to external parameters" (Toulouse 1983). It turns out that some of these features are important in other optimization problems also, whence the wide interest in spin glass.

3.2 *A rapid overview*

Theoretical studies on spin glass have yielded a rich harvest, and many landmarks can be identified. First there is the paper by Edwards and Anderson (1975) which introduced not only a convenient (though idealized) model but also several key concepts, including a suitable order parameter known as the Edwards–Anderson parameter. The next major milestone is the attempt of Sherrington and Kirkpatrick (1975, abbreviated as SK; see also Kirkpatrick and Sherrington 1978) to construct an exact mean field theory. Though some of the results of SK are unacceptable, their work nevertheless exposed many subtle questions leading later to important discoveries. Thouless *et al* (1977, abbreviated as TAP) found a way of handling the mean field theory which avoided the pitfalls encountered by SK. Among other things, they demonstrated the existence of many solutions, implying the existence of many equivalent or near-equivalent spin configurations. Meanwhile, there was some uncertainty concerning the nature of the order parameter until Parisi (1979, 1980a, b, c, 1983) cleared the fog. He not only formalized a new concept called *replica symmetry breaking* (analogous to the breaking of continuous symmetry during the familiar second-order phase transitions), but also provided an *ansatz* for implementing the breaking. Finally, Mezard *et al* (1984a, b) discovered that the family of spin glass solutions had a hierarchy of their own, implying in turn a very rugged free energy terrain (recall figure 1(c)). We shall now discuss these developments in slightly greater detail.

3.3 *The Edwards–Anderson model*

The Edwards–Anderson (EA) model is described by the Hamiltonian

$$\mathscr{H} = \sum_{\langle ij \rangle} J_{ij} S_i \cdot S_j \tag{6}$$

where the S_i's are classical vector spins located at the sites i of a *regular* lattice and the sum $\langle ij \rangle$ is over the nearest-neighbours. The features illustrated in figure 7 are comprehensively accommodated by making the exchange interaction J_{ij} into a random variable with the distribution

$$p(J_{ij}) = (1/2\pi J^2)^{1/2} \exp(-J_{ij}^2/2J^2). \tag{7}$$

The disorder is said to be *quenched* by which is meant that the values of J_{ij} are frozen. Notice the idealization introduced by the theorist while modelling the spin glass. Interestingly, such models yield quite meaningful results.

A quenched system is one in which some degrees of freedom, say $\{A_{ij}\}$, are frozen while others $\{B_{ij}\}$ are in thermal equilibrium. On the other hand, situations can arise where the $\{A_{ij}\}$ are not frozen but *annealed.* In this case (which is the usual one), the free energy is given by

$$-\beta F = \ln [\mathrm{Tr}_{\{A,B\}} \exp(-\beta \mathscr{H}\{A\}, \{B\})] \tag{8}$$

i.e., the trace is over *both* sets of variables. Referring to figure 7, if the magnetic atoms can diffuse, the exchange interactions will exhibit *thermal fluctuations*, and one would evaluate the free energy as in (8). However, if diffusion is negligible, exchange disorder must be regarded as quenched which is the assumption of the model defined by (6) and (7). Edwards and Anderson indicate how F must be calculated in such a case.

Consider a system obeying (6) and (7). Here the exchange constants J_{ij} are frozen, and the set $\mathscr{J} = \{J_{ij}\}$ represents one *instance* of the possibilities available for the quenched variables. The free energy $\mathscr{F}(\mathscr{J})$ of this sample is given by

$$\mathscr{F}(\mathscr{J}) = -k_B T \ln Z(\mathscr{J}). \tag{9}$$

The observed free energy is then given by

$$F \equiv [\mathscr{F}(\mathscr{J})]_{\mathrm{Av}} = \int d\mathscr{J} P(\mathscr{J}) \mathscr{F}(\mathscr{J}) \tag{10}$$

where $P(\mathscr{J})$ denotes the probability for the occurrence of the disorder pattern $\mathscr{J}$. The averaging in (10) is referred to as *configuration averaging*, and its significance will be discussed later.

Reference must now be made to the famous Edward–Anderson order parameter q_{EA}. The concept is simple and has already been schematically illustrated in figure 6. More formally,

$$q_{\mathrm{EA}} = \lim_{t\to\infty} \lim_{N\to\infty} q(t) \tag{11a}$$

where $q(t)$ denotes a time average performed over an observation time $\mathscr{T}_{\mathrm{obs}}$ and is given by

$$q(t) = \left[(1/\mathscr{T}_{\mathrm{obs}}) \int_0^{\mathscr{T}_{\mathrm{obs}}} \mathrm{d}t'\, S_i(t') S_i(t+t') \right]_{\mathrm{Av}} . \tag{11b}$$

As in (10), $[\ ]_{\mathrm{Av}}$ in the above equation implies a configuration averaging. Alternatively, and this is more common, one defines

$$q_{\mathrm{EA}} = [\langle S_i \rangle_T^2]_{\mathrm{Av}} \tag{12}$$

where $\langle\ \rangle_T$ is the traditional thermal average taken here in a particular configuration $\mathscr{J}$.

EA have pointed out that the physical picture underlying (11) and (12) have parallels in polymer science. When a solution of very long molecules becomes dense, there comes a density at which the mobility of a molecule falls essentially to zero and the system gels. Such a molecule will then appear as the *same* random coil at various instants of time i.e., there is a freezing of the *form* as well as the *orientation*.

The paper of Edwards and Anderson is also noteworthy for the introduction of the *replica trick*, about which more will be said in the following subsection. Here we just note that using their model EA were able to show that a cusp occurs in the curves for the specific heat and the susceptibility, and that $q_{\mathrm{EA}} \neq 0$ for temperatures lower than the cusp temperature.

3.4 *Mean field theory of Sherrington and Kirkpatrick*

The goal of SK was to construct a mean field theory for spin glass. Mean field theories enjoy a privileged place in condensed matter physics, thanks to their immense utility first exhibited in the field of ferromagnetism. They are one of a large class of cluster approximations in which the interactions of particles within a cluster are treated exactly while those involving a particle outside are treated in an average fashion. In the so-called molecular field approximation to the Ising model of ferromagnetism, the cluster becomes single spin. The outcome is the well-known self-consistency equation

$$m = \tanh(\beta J z m) \tag{13}$$

where m is the magnetization (per spin), J the (nearest neighbour) exchange integral and z the number of nearest neighbours.

Stanley (1971) has shown that the mean field theory of ferromagnetism becomes exact in the limit of infinite range i.e., every spin interacts with every other spin instead of with near neighbours alone. Guided by this, SK proposed that the mean field theory of spin glass should be the exact solution of the infinite range model similar to (6). Their Hamiltonian is

$$\mathscr{H} = -\sum_{(ij)} J_{ij} S_i S_j - h\sum_i S_i \tag{14}$$

where the spins are of the Ising type i.e., take the values ± 1, and the sum (ij) is over *all* neighbours. The quantity h denotes a uniform external field. As before,

$$p(J_{ij}) = (1/2\pi J^2)^{1/2} \exp[-(J_{ij} - J_0)^2/2J^2], \tag{15}$$

a small difference being that an offset from zero has been added for generality. In passing we note that with an infinite-range model, questions related to the structure of the lattice and its dimensionality become irrelevant. The assumption of an infinite range is clearly unphysical but since it leads to sensible results for a ferromagnet, there is a case for trying it out for the spin glass as well. We now digress to discuss the replica trick.

Consider the problem of calculating $f(\beta)$ the free energy per particle of a many-particle system. The rule specified by equilibrium statistical mechanics is

$$f(\beta) = \lim_{N\to\infty} F_N(\beta)/N. \tag{16}$$

Problems arise, however, if the system has randomness. As discussed earlier, in a system with quenched random interactions, $f(\beta)$ has to be obtained by averaging $F_N(\beta)$ over the distribution $P(\mathscr{J})$. This average must be computed *after* taking the logarithm of the partition function, but *before* taking the thermodynamic limit i.e.,

$$-\beta f(\beta) = \lim_{N\to\infty} (1/N)[\ln Z(\beta)]_{\mathrm{Av}} \tag{17}$$

where, for convenience, we write Z instead of Z_N. Averaging the logarithm of the partition function is no easy task but there is a way out which is to exploit the identity

$$\ln x = \lim_{n\to 0} (x^n - 1)/n. \tag{18}$$

The identity is exploited by imagining first that n identical copies or replicas are made of the given system (making in all $(n+1)$ systems i.e., one original plus n copies). As will be seen shortly, this proliferation aids the evaluation of the configuration average. Once this averging is performed there is no further need for the replicas, which are then "disposed off" by taking the limit $n\to 0$. Turning to the details, we first have

$$\begin{aligned}[\ln Z(\beta)]_{\mathrm{Av}} &= \lim_{n\to 0} \{[Z^n]_{\mathrm{Av}} - 1\}/n \\ &= \lim_{n\to 0} \left\{-1 + \int \mathrm{d}\mathscr{J} P(\mathscr{J}) \prod_{\alpha=1}^{n} Z_\alpha\right\} \Big/ n.\end{aligned} \tag{19}$$

where Z_α is the partition function of the αth replica. $P(\mathscr{J})$ is the same as in (10). Now we can write

$$\prod_{\alpha=1}^{n} Z_\alpha = \mathrm{Tr}_{\mathrm{RS}}\left[\exp\left(-\beta \sum_{\alpha=1}^{n} \mathscr{H}^\alpha\right)\right] \tag{20}$$

where $\mathscr{H}^1, \mathscr{H}^2, \ldots\ldots \mathscr{H}^n$, denote the Hamiltonians of n replicas of the system. The trace RS in (20) is over all the nN (replica) spins associated with the n replicas. Blandin *et al* (1980) have noted that from a physical point of view, two replicas α and β can also be understood as the *same* system at two times t_1 and t_2 with $|t_1 - t_2| \to \infty$.

Upon introducing (20) in (19), the configuration averaging is readily performed since the Gaussian integrals can be evaluated, and one obtains

$$[Z^n]_{\mathrm{Av}} = \mathrm{Tr}_{\mathrm{RS}} \exp\{-\beta \mathscr{H}_{\mathrm{eff}}(n)\} \tag{21}$$

where $\mathscr{H}_{\mathrm{eff}}(n)$ is the *effective* Hamiltonian of the n-replica system and is given by

$$\beta \mathscr{H}_{\mathrm{eff}}(n) = -\sum_{ij}\left\{(\beta J_0/2)\sum_{\alpha} S_i^\alpha S_j^\alpha + (\beta J/2)^2 \times \sum_{\alpha,\beta} S_i^\alpha S_j^\alpha S_i^\beta S_j^\beta\right\}. \tag{22}$$

Using the results of the above two equations, the $n \to 0$ limit in (19) can be implemented leading to $[\ln Z(\beta)]_{\mathrm{Av}}$ from which $f(\beta)$ is then obtained via (17).

Let us pause and absorb the import of the above. We started with the nasty problem of averaging the logarithm of the partition function. We then sought a way out via the identity (18). Raising the partition function to the power n brought in n replicas of the SK Hamiltonian but this proliferation was worth it since the messy $\mathscr{J}$ integration could be got rid off, leaving us with the trace over nN spins of an effective Hamiltonian $\mathscr{H}_{\mathrm{eff}}$ defined in the space of variables $\{S_i^\alpha\}$ of all the n replicas of the system. Observe that $\mathscr{H}_{\mathrm{eff}}$ has no elements of randomness left in it and is translationally invariant. Also, the trace to be performed in (21) represents an annealed average.

One would have thought that the use of (18) in (17) would have swept away all the difficulties; not quite! Firstly, (17) and (20) together imply that the $n \to 0$ limit is taken *before* the $N \to \infty$ limit, whereas one must really perform the $N \to \infty$ limit earlier in order to properly evaluate Z. This interchange of limits has attracted comments (e.g., TAP, van Hemmen and Palmer 1979) but now seems to be an acceptable procedure. More serious is the problem of going from a positive integer n to a real n in the neighbourhood of $n = 0$, since the replicated Hamiltonian is not well defined under these circumstnces. This is a sufficiently important issue for us to return to it later.

Resuming our description of the work of SK, the step leading to a mean field theory is implemented by replacing as usual, $S_i^\alpha S_j^\alpha$ by $S_i^\alpha \langle S_j^\alpha \rangle$ and $S_i^\alpha S_j^\alpha S_i^\beta S_j^\beta$ by $S_i^\alpha S_i^\beta \langle S_j^\alpha S_j^\beta \rangle$. The averages $\langle\ \rangle$ are now over the states of the effective Hamiltonian, and there are no problems with configuration averaging any longer. The order parameters

$$m_\alpha \equiv \langle S_j^\alpha \rangle \text{ and } q_{\alpha\beta} = \langle S_i^\alpha S_j^\beta \rangle \quad \alpha \neq \beta \tag{23}$$

are to be determined self-consistently. At this stage, SK make the approximation $m_\alpha = M$ and $q_{\alpha\beta} = q$ (i.e., there is no dependence on the replica indices). This then leads to a relatively simple expression for βf, with attendant self-consistency

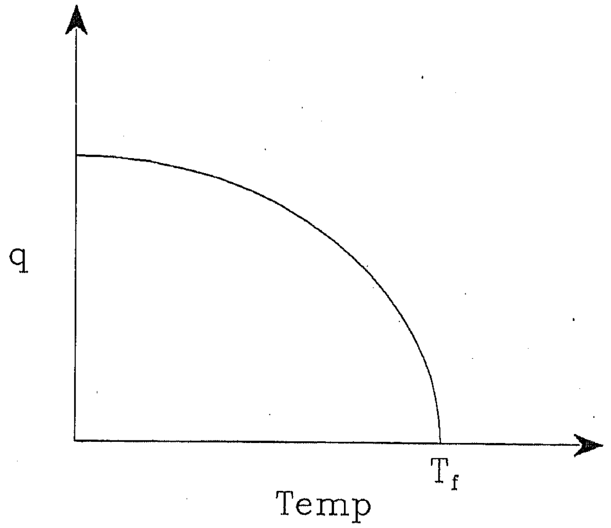


Figure 8. Temperature variation of the spin glass order parameter q (schematic).

requirements on M and q. From a practical point of view, finding βf involved minimizing a (nxn) matrix derived from the effective Hamiltonian.

SK calculated the temperature variation of q, with features as in figure 8. No doubt this conforms to expectations, but the difficulty with the SK solution is that it leads to negative entropy which is unacceptable. The root cause is that the SK solution is *unstable*. The question of how to rectify this problem will be taken up in § 3.7.

Notwithstanding the unphysical nature of the SK solution, one firm conclusion to emerge from their work (which, incidentally, also included numerical simulations), is that in the mean field theory, there is a phase transition to a spin glass phase at a non-vanishing temperature T_f, the free energy terrain of this phase consisting of deep valleys separated by barriers with high activation energies. In the thermodynamic limit, the phase transition is sharp and the barriers become infinite. Also, their number increases as a power of N.

3.5 *Mean field theory of TAP*

Replicas provide a convenient means of performing the average over disorder. Thouless *et al* (1977) suggested that one defer this average to the end, and, following tradition, first wrote down the local-site mean field equations for a given disorder pattern $\mathscr{J}$. Now for the Ising ferromagnet (with no disorder), the equation is of the form

$$m_i = \tanh\left[\beta\left(\sum_{j\neq i} J_{ij}\right)m_i\right]. \qquad (24)$$

Of course m_i is the same for all i. More general ordered states (e.g. anti-ferromagnetic) can be allowed for by having

$$m_i = \tanh\left[\beta\sum_j J_{ij}m_j\right]. \qquad (25)$$

In the case of spin glass TAP recognized that some further modifications are necessary, after incorporating which they deduced

$$m_i = \tan\mathrm{h}\left[\beta\sum_j J_{ij}m_j - \beta\sum_j J_{ij}^2(1-m_j^2)m_i\right]. \tag{26}$$

It is useful to look upon (26) as the stationarity conditions

$$\partial F\{m\}/\partial m_j = 0 \tag{27}$$

where $F\{m\}$ is the TAP free energy (for a given configuration $\mathscr{J}$). Observe that (26), (27) finally lead to a set $\{m\}$ of N site magnetizations which collectively describe the spin glass state. The message is that to describe the order parameter, a single number like q (as in the SK model) would not do. This again is a point to which we shall return later.

Equations (26) and (27) are not particularly easy to handle since J_{ij} is a random matrix. However, both near the transition temperature and near $T=0$, some simplifications are possible. With the aid of these and some numerical work, TAP were able to assert that physically meaningful results *could* be obtained from mean field theory and that the earlier conclusions of EA and SK about a sharp transition leading to a totally random frozen state were completely valid.

The number of solutions (26) is enormous, being of the order $\exp[\alpha(T)N]$ where $\alpha(T)$ varies from about 0·2 at $T=0$ (Bray and Moore 1980; de Dominics *et al* 1980; Tanaka and Edwards 1980) to zero at $T=T_f$. Notionally the various solutions should all have the same free energy f_0 per spin (to leading order in N) whence one speaks of ground state degeneracy. Now the statistical expectation value of a quantity is the average of the various values that such a quantity takes in all possible situations. Guided by the consideration just given, the first suggestion was to calculate average over all possible TAP states by assigning equal weights to them. This choice called *white averaging*, led to a worse disaster than the pre-Parisi replica solution. It was then realized that just because all the free energies f_s tend to the same value f_0 (to leading order in N that is) it does not necessarily mean that all the states s have the same weight p_s. On the contrary, there are $O(1/N)$ differences in the f_s which cause the weight

$$p_s = \exp[-\beta N f_s]\Big/\sum_t \exp[-\beta N f_t] \tag{28}$$

to vary by a significant amount (Bray *et al* 1984, 1986). It is the totality of all states whose number varies exponentially with N. However, if f_c denotes the minimum free energy, the numerous states with $f \gg f_c$ exert hardly any influence, and the states of interest are those with $f \sim f_c$. At $T=0$, the state with $f=f_c$ would represent the global minimum, though in a notional sense perhaps. This point will be amplified with a figure shortly.

3.6 *The phase diagram*

We can now sketch the phase diagram, and pause to absorb its significance. The control parameters are the temperature T, the external magnetic field h and those which characterize the distribution (15) namely, J_0 and J. In terms of these, the phase

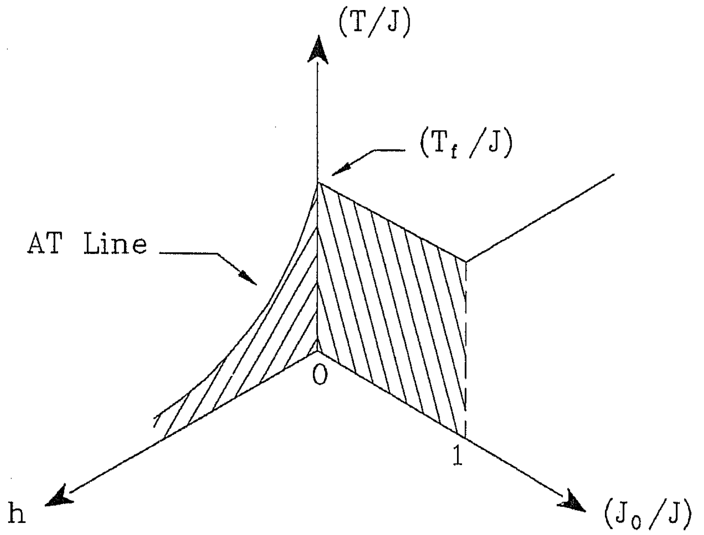


Figure 9. Schematic phase diagram for spin glass. SK explored along the T-axis and discovered a transition at T_f. However, their solution is unstable below T_f. In fact, there is instability in the shaded region, unless replica symmetry is broken as discussed in § 3.8.

diagram may be comprehensively depicted as in figure 9. Considering first the temperature axis, SK showed that in the mean field theory one should expect a phase transition to the spin glass phase below T_f. As noted earlier, their solution is unstable; indeed this instability extends both to the $(J_0 - T)$ as well as the $(h - T)$ planes, and is shown by the shaded region. The line marked AT is that predicted by de Almeida and Thouless (1978), being the boundary separating the stable region from the unstable in the $(h - T)$ plane. The cure for the instability of the SK solution is our next concern.

3.7 *The Parisi solution*

Before discussing how Parisi cured the problem of instability, it is useful to recall first that in a conventional phase transition, symmetry is said to be broken if the thermodynamic state lacks the symmetry possessed by the free energy expansion (in terms of the order parameter). Now while seeking a solution, SK assumed that $q_{\alpha\beta} = q$, independent of the indices α and β. Effectively this implies that the mean field solution is symmetric under a permutation of the replica indices, even as the replica Hamiltonian itself is. This is certainly true for integer n but there is no guarantee that it also holds when $n \rightarrow 0$.

The instability of the SK solution prompted a search for others which *broke* replica symmetry, culminating in a series of papers by Parisi (1979, 1980a, b, c, 1983) which settled the issue.

Parisi realized that a proper description of spin glass requires not one but an *infinity* (!) of order parameters, in fact with a hierarchical structure (which will be commented upon in § 3.9). Parisi builds up the required hierarchy as follows: Start with a $(n \times n)$ matrix (temporarily and for convenience, we shall write m_0 instead of n). Divide the $(m_0 \times m_0)$ matrix into blocks of size $(m_1 \times m_1)$ as in figure 10. Assign a value q_0 to all matrix elements except those in the diagonal boxes. (If this value is also assigned to the elements of the diagonal boxes, then one would be back to what SK did.) The diagonal boxes are next divided into boxes of size $(m_2 \times m_2)$ (see figure 10(b)), and a

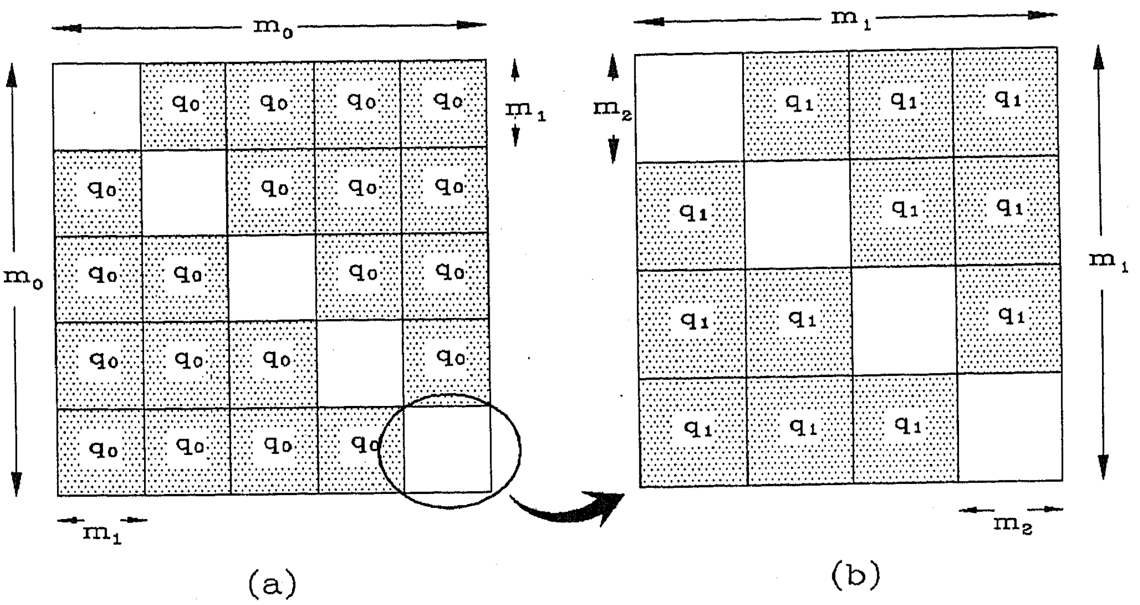


Figure 10. Hierarchical matrix partitioning scheme of Parisi. For explanations, see text.

value q_1 is assigned everywhere except again in the diagonal boxes. This process is repeated say $k+1$ times such that

$$n = m_0 \geqslant m_1 \geqslant m_2 \cdots \geqslant m_k \geqslant 1.$$

After the last partitioning, zero is entered in all the diagonal boxes. The order parameter is now described by the set $\{q_0 = q(m_0),\ q_1 = q(m_1),\ q_2 = q(m_2), \cdots\}$.

So far everything is simple and clear. Parisi now lets $n \to 0$ and simultaneously $k \to \infty$. This is hard to visualize for it essentially means collapsing the starting $(n \times n)$ matrix into a (0×0) matrix, while at the same time endowing it with infinite structure! Mercifully, what emerges for ordinary mortals is that the order parameter set referred to above now becomes a continuous function $q(x)$ defined in the interval $0 \leqslant x \leqslant 1$. (Caution: x here should not be confused with a co-ordinate variable; rather, it is descended from m_k.) If replica symmetry is unbroken, $q(x)$ is a constant; otherwise it has a form which will be sketched presently.

The general solution for $q(x)$ is not known but close to T_f one can obtain a Landau-type expansion

$$-\beta f = \beta f_0 + 1/2 \int_0^1 \mathrm{d}x \left\{ |\theta| q^2(x) + q^4(x)/6 - x q^3(x)/3 - q(x) \int_0^x \mathrm{d}y\, q^2(y) \right\} \tag{29}$$

where $\theta = (T - T_f)/T_f$. Figure 11 offers a perspective on $q(x)$ without and with replica-symmetry breaking. Since q can assume a range of values, it becomes meaningful to talk of a probability distribution for q namely $\mathscr{P}(q)$, the form of which is also sketched in the above figure.

3.8 *Some implications of the Parisi solution*

The Parisi solution came like breath of fresh air and spurred many investigations which collectively have provided much clarification. We present below a medley of the ideas which have emerged.

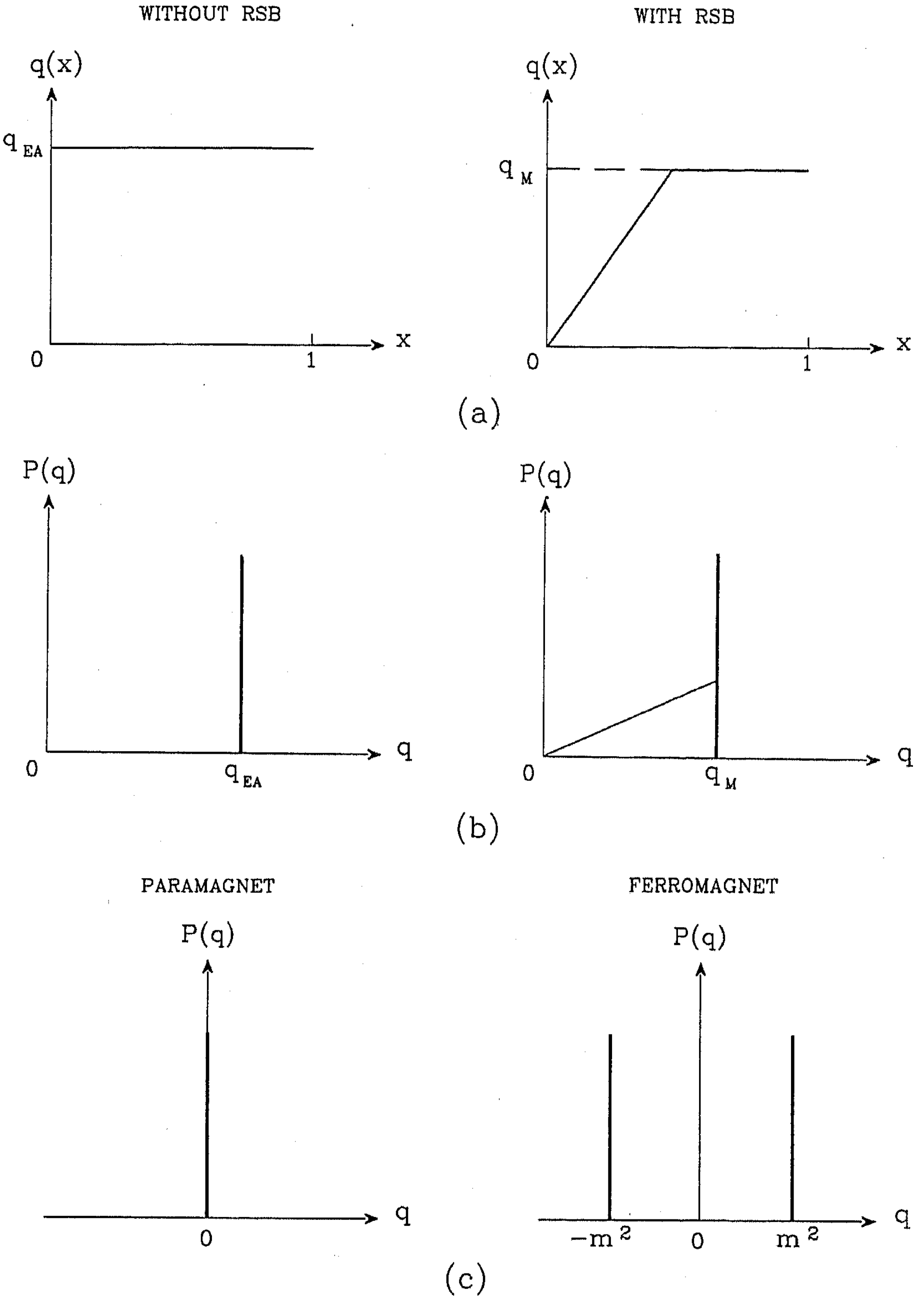


Figure 11. (a) Shows the behaviour of $q(x)$ without and with replica symmetry breaking. The corresponding distribution functions $\mathscr{P}(q)$ are shown in (b). For fixing our ideas, $\mathscr{P}(q)$ for a para and a ferromagnet is shown in (c).

3.8.1 *An infinity of order parameters*: In the usual order-disorder transformation, one has a single number (like the magnetization M for example) to characterize the (perfectly) ordered state. Parisi's work shows that one needs a whole function $q(x)$ corresponding to what he refers to as an "infinite number of order parameters" (Parisi 1979). The function itself is deduced by studying the extrema of an expression like (29). The message about an infinity of order parameter is also implicitly contained in TAP's set $\{m_i\}$. Going back to figure 1, the curve in 1(c) is really a hypersurface in the space of all TAP states (of dimensionality $\sim \exp N$). What is shown there is highly schematic, the main purpose being to draw attention to the existence of a large number of valleys with varying depths (as mentioned while discussing the TAP solutions).

3.8.2 *Connection between TAP and Parisi solutions*: It would have been noticed that TAP and Parisi follow somewhat different routes, although towards the same global objective. Whereas TAP deal in detail with the mean field equations for the $\mathscr{J}$ relegating configuration averaging for later consideration, Parisi who takes off from SK already has the configuration averaging behind him, being left only with the averaging over replicas (which he then handles in an ingenious manner). The linkages between the TAP and the Parisi approaches are now available.

Consider a sample corresponding to a given $\mathscr{J}$. Let s, t denote two TAP solutions i.e. s and t correspond to two valleys (or local minima). Within a valley the system has some freedom for thermal fluctuations, and we denote the thermal average of an observable A in the valley s as $\langle A \rangle_T^s$; it is a partial average, being defined by

$$\langle A \rangle_T^s = (1/Z_s) \sum_{\lambda \in s} A_\lambda \exp(-\beta N f_\lambda). \tag{30}$$

The Gibbs average $\langle A \rangle_T$ which we normally are used to i.e.,

$$\langle A \rangle_T = (1/Z) \sum_{\lambda} A_\lambda \exp(-\beta N f_\lambda) \tag{31}$$

can be written as

$$\langle A \rangle_T = \sum_{s} p_s \langle A \rangle_T^s \tag{32}$$

where

$$p_s = \exp(-\beta N f_s) \Big/ \sum_{t} \exp(-\beta N f_t). \tag{33}$$

With these definitions, the valley overlap may be expressed as

$$q_{st} = (1/N) \sum_{i=1}^{N} \langle S_i \rangle_T^s \langle S_i \rangle_T^t, \tag{34}$$

and the overlap probability function $\mathscr{P}_{\mathscr{J}}(q)$ as

$$\mathscr{P}_{\mathscr{J}}(q) = \sum_{s} \sum_{t} p_s p_t \delta(q - q_{st}). \tag{35}$$

$\mathscr{P}_{\mathscr{J}}(q)$ is sample dependent and can fluctuate with $\mathscr{J}$, as a result of which there is a probability distribution for the probability distribution! This calls for one more round of averaging, leading to

$$\mathscr{P}(q) \equiv [\mathscr{P}_{\mathscr{J}}(q)]_{\text{Av}}. \tag{36}$$

Let q_{max} and q_{min} denote the maximum and minimum values of q (at a given temperature and a given magnetic field). Obviously,

$$-1 \leqslant q_{\text{min}} \leqslant q_{\text{max}} \leqslant 1.$$

Define now

$$x(q) = \int_{-1}^{q} dq' \mathscr{P}(q') \tag{37}$$

which is a cumulative distribution giving the probability of having all overlaps from -1 up to a stipulated value q. With $x(q)$ available, one can also define an inverse function $q(x)$. Interestingly but perhaps not surprisingly, the $q(x)$ so obtained by starting with TAP solutions is the same as the Parisi function $q(x)$ introduced in § 3.6, although the latter was arrived at via the replica route.

Table 1 shows some equivalences, while figure 12 shows the form of $\mathscr{P}(q)$ as predicted and as observed in numerical simulations.

3.8.3 *Self-averaging*: Anderson (1978) has remarked, "No experiment is ever done on an ensemble of samples." Very true indeed. An experimentalist might repeat his experiment with several samples to confirm his result but not for averaging in the same sense a theorist does (as in (10), for example). In the spin glass case, this might at first sight be puzzling, considering that a given (experimental) sample might correspond to *one* particular realization $\mathscr{J}$ of the many patterns of random exchange interactions possible. However, assuming that the experimentalist works with a sample that is effectively infinite, such a sample could be visualized as divided into a number of subvolumes each of which is representative of a particular pattern of quenched disorder. Thus it is that one assumes that the properties of *one large sample* differs negligibly from a *configuration average over an ensemble* of systems with different instances of quenched disorder. When such an equivalence obtains, the physical quantity is said to be *self-averaging*. Note that in the single large sample, the quantities J_{ij} are quenched and not annealed.

Table 1. Some quantities as defined in the TAP approach and as defined in the replica approach.

TAP	Replica
$q = [\langle S_i \rangle_T^2]_{\mathrm{Av}}$	$q = \lim\limits_{n\to 0} \frac{1}{n(n-1)} \sum\limits_{\alpha\neq\beta} q_{\alpha\beta}$
$q^{(2)} = \frac{1}{N} \sum\limits_{i,j} \left[\left(\sum\limits_l p_l \langle S_i \rangle_T^l \langle S_j \rangle_T^l \right)^2 \right]$	$q^{(2)} = \lim\limits_{n\to 0} \frac{1}{n(n-1)} \sum\limits_{\alpha\neq\beta} q_{\alpha\beta}^2$
$q_{\mathrm{EA}} = \left[\sum\limits_l p_l (\langle S_i \rangle_T^l)^2 \right]_{\mathrm{Av}}$	$q_{\mathrm{EA}} = \lim\limits_{n\to 0} \lim\limits_{\alpha\to\beta} q_{\alpha\beta}$
$[\langle S_i S_j \rangle_T^2]_{\mathrm{Av}}$	$\lim\limits_{n\to 0} \frac{1}{n(n-1)} \sum\limits_{\alpha,\beta} \langle S_i^\alpha S_j^\alpha S_i^\beta S_j^\beta \rangle$
$\mathscr{P}(q) = \left[\sum\limits_{lm} p_l p_m \delta(q - q_{lm}) \right]_{\mathrm{Av}}$	$\mathscr{P}(q) = \lim\limits_{n\to 0} \frac{1}{n(n-1)} \sum\limits_{\alpha\neq\beta} \delta(q - q_{\alpha\beta})$

$$q = \int_0^1 \mathrm{d}x\, q(x)$$

$$q_{\mathrm{EA}} = q(x = 1)$$

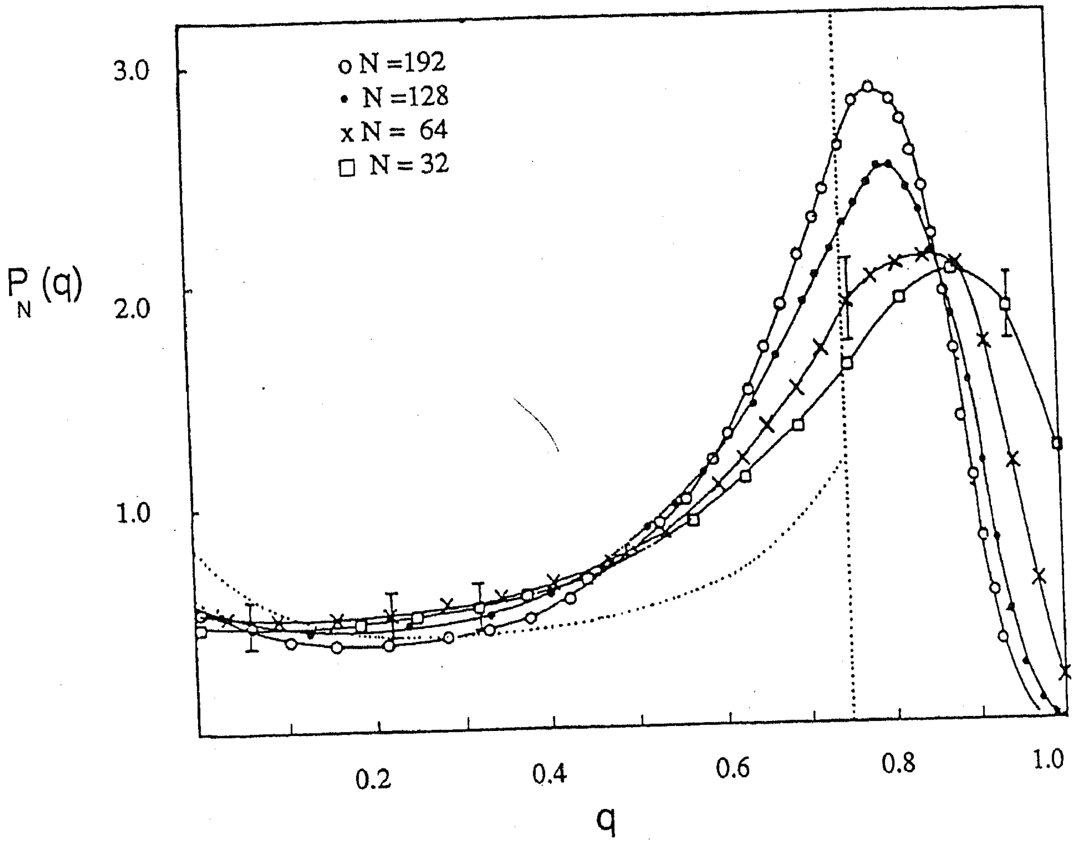


Figure 12. $\mathscr{P}(q)$ for the SK model as determined from numerical studies. The dotted line is the prediction of theory, and consists of a delta function and a continuous part (after Young 1983).

Generally speaking, self-averaging applies to extensive variables associated with systems in which the interactions are short ranged. Prima facie therefore, one would not expect any quantity in the SK model to be self-averaging since the interactions are of infinite range. However, it turns out that quantities which are self-averaging in the short range model are also self-averaging in the SK model.

The existence or the absence of self-averaging can be tested as follows: Let O denote a physical quantity, $\langle O\rangle_T$ its thermal average corresponding to the configuration $\mathscr{J}$, and $[\langle O\rangle_T]_{\mathrm{Av}}$ its configuration average. If O is self-averaging, then the variance

$$\Delta O \equiv [(\langle O\rangle_T)^2]_{\mathrm{Av}} - ([\langle O\rangle_T]_{\mathrm{Av}})^2 \qquad (38)$$

is zero. Physically observable quantities like the magnetization m and the internal energy E satisfy (38) and are thus self-averaging. One of the surprises in this field has been that there exist quantities which are *not* self-averaging; examples include q and $\mathscr{P}_{\mathscr{J}}(q)$ (see also figure 13). This is a point requiring comment.

Young *et al* (1984) have noted that quantities which are not self-averaging are closely linked to overlap. Physically this seems to imply that for some properties, not only are the valleys in which the system is resident important, but also their proximity to other valleys. While the *overall* valley structure might appear to change marginally from one $\mathscr{J}$ to another, overlap structure shows considerable variations (as reflected in the variations of $\mathscr{P}_{\mathscr{J}}$). And where overlap is important (as in the case of q), the

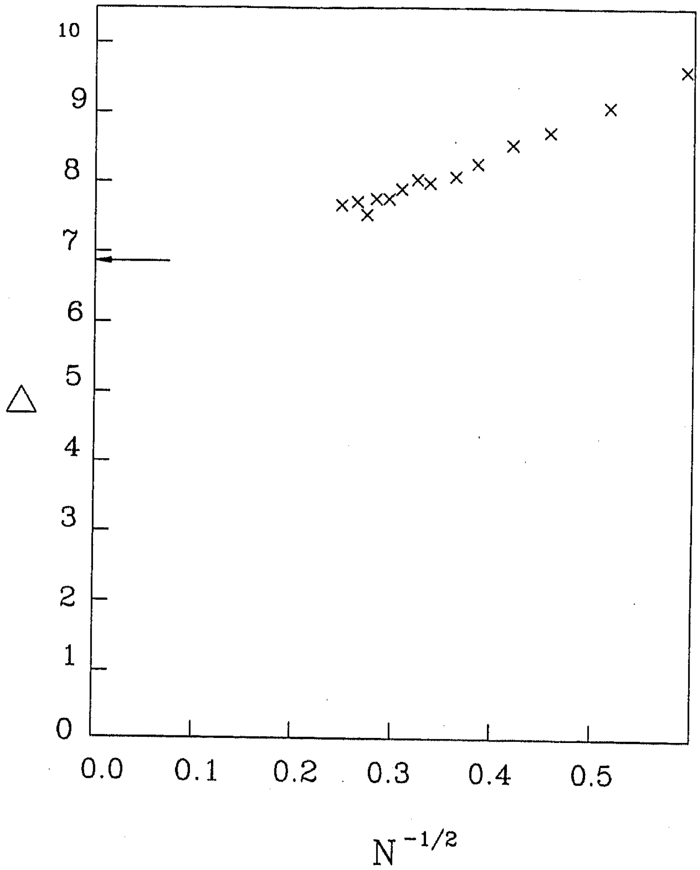


Figure 13. Variance Δ (as defined in eq. (38)) for the overlap parameter q, plotted as a function of sample size N. The arrow shows predictions from Parisi's theory (after Young *et al* 1984).

corresponding property ceases to be self-averaging. So far no example of this is available from laboratory experiments but some based on computer simulation have been reported.

3.8.4 *Freezing and ergodicity breaking*: Any description of freezing depends crucially on the observation time scale $\mathscr{T}_{\text{obs}}$. Figure 14 amplifies the idea (Palmer 1982). Here the horizontal lines represent the time axis and the shaded patches the time scales of the relaxation processes. In figure 14(a), $\mathscr{T}_{\text{obs}} \gg$ all relaxation times, and the system may be said to be in equilibrium. This is the classic, text-book situation where averages are obtained using the Gibb's prescription, and the system wanders freely over the whole of phase space "visiting every hamlet" (van Hemmen 1983). Such an ideal situation does not always obtain, for example in a phase transition. Figure 14(b) represents the scenario for an Ising ferromagnet which has condensed into one of the two possible states of ordering, say $(M+)$. On reasonable times scales of observation, all the relaxations in the $(M+)$ state would have occurred and the system may be considered to be in partial equilibrium with its wanderings confined to the pocket $\Gamma(+)$ of the phase space Γ. However, relaxations associated with flips $+M \rightarrow -M$ and vice versa are still possible, but to observe these $\mathscr{T}_{\text{obs}}$ would have to be made very much larger. If this is done, one would observe the system randomly jumping back and forth between the ± states. The system can be regarded as being in full

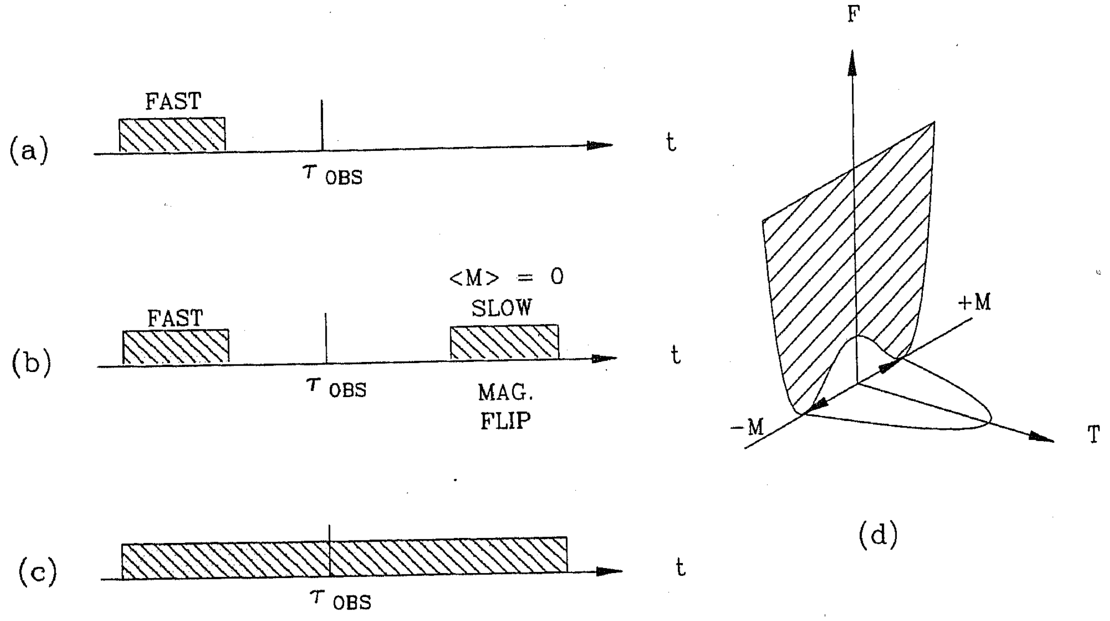


Figure 14. Schematic illustration of the relaxation times for various systems in relation to the observation time τ_{obs}. The time axis is supposed to be on a logarithmic scale. In (a) the relaxation is fast, and the system is in equilibrium. An Ising ferromagnet is illustrated in (b). The system is ordered, and on the scale τ_{obs} appears to remain in a fixed state of magnetization. However, magnetization flips are possible [see also (d)], though on a very much longer time scale. Case (c) pertains to a spin glass.

thermodynamic equilibrium but now $\langle M \rangle = 0$ which is uninteresting. Fortunately, Nature cooperates by making $\mathscr{T}_{FLIP}$ astronomically large whence a system condensing into a particular ordered state will remain in that state practically for ever (at $T = 0$ certainly, and more reasonably also at non-zero temperatures below the transition temperature).

The spin glass situation is more complicated, being rather like in figure 14(c) where there is a continuum of relaxation times. Whatever freezing one observes is with respect to processes *slower* than $\mathscr{T}_{\text{obs}}$.

A physical system is said to *break ergodicity* when its behaviour is qualitatively and quantitatively different from that predicted by equilibrium statistical mechanics. This happens in situations like in figures 14(b) and (c) where $\mathscr{T}_{\text{obs}}$ is less than some of the relaxation times occurring in the system. Under these circumstances, the phase space Γ becomes divided into disjoint components Γ^α with $\Gamma = \cup_\alpha \Gamma^\alpha$. Depending on the initial conditions, the system gets trapped into *one of these components.* This is a very important feature of spin glass.

How does one calculate the properties of such a system? Within Γ^α there is no problem since the standard rules of statistical mechanics apply, with the restriction that they are applied in the domain of Γ^α. Averages $\langle \cdots \rangle_T^\alpha$ computed in this manner [cf. equation (30)] are sometimes referred to as *restricted* averages. The problem is that one does not know which Γ^α the system is trapped in, a difficulty which is particularly acute in the case of spin glass. Something more than a mere restricted average is therefore called for. One approach is to calculate the canonical average (of the observable A) by the standard rule

$$A_c = \langle A(X) \rangle_T^\Gamma = \mathrm{Tr}[X \in \Gamma] A(X) \exp\{-\beta \mathscr{H}(X)\}/Z \tag{39}$$

where $\mathscr{H}$ is the Hamiltonian, and the trace is over the configuration variable X. In the case of other quantities (like specific heat, for example) A_c is computed from the partition function and its appropriate derivatives. Instead of A_c, one could also compute a component average $\bar{A}$ defined by

$$\bar{A} = \sum_{\alpha} p_{\alpha} A_{\alpha} \tag{40}$$

where

$$A_{\alpha} = \langle A(X) \rangle_T^{\alpha} = \mathrm{Tr}[X \in \Gamma^{\alpha}] A(X) \exp\{-\beta \mathscr{H}(X)\}/Z_{\alpha}. \tag{41}$$

The real system is described by $p_{\alpha} = 1$ for some $\alpha = \alpha_0$ (say) and $p_{\alpha} = 0$ for all other values of α. But we do not know α_0 which is precisely why all the fuss about finding some kind of an average. Given our ignorance about the initial state, it is natural in the spirit of statistical mechanics to suppose that

$$p_{\alpha} = exp(-\beta N f_{\alpha})/Z = Z^{\alpha}/Z. \tag{42}$$

With these definitions, the following are some of the results which ensue (Palmer 1983):

(i) $A_c = \bar{A}$ for observables like magnetization, interval energy etc.
(ii) $\bar{F} = F_c + TI$ and $\bar{S} = S_c - I$ where

$$I = -k_B \sum_{\alpha} p_{\alpha} \ln p_{\alpha} \tag{43}$$

is the intercomponent entropy and is referred to as *complexity*. Physically it denotes the additional information needed to specify a particular state, given the *a priori* probabilities.

There are similar rules for the specific heat, susceptibility etc.

Each component α is associated to a TAP state. Sometimes, the phrase *pure state* is also used; thus, $\bar{A}$ is a mixture of pure states. (Thermodynamists may perhaps be more comfortable in referring to these pure states as phases). Besides the association referred to above, there are also some finer details to be taken note of. Figure 15 shows qualitative plots of the free energy hypersurface as T is lowerd. Below T_f, one has a rugged terrain with deep valleys. As T decreases, the valley structure becomes richer and more rugged -essentially with the development of "valleys within valleys within valleys ......" (Binder and Young 1986). Averages $\langle \cdots \rangle_T^{\alpha}$ are associated to the wanderings within limited region Γ^{α} of spin configuration space: An example of such confinement has been given earlier in figure 4.

Valley bifurcation as T is lowered can be given a hierarchical representation as in figure 15(b). A word of caution: Although the (inverted) tree in figure 15(b) is shown as having one "trunk" just below T_f, there should in reality be an infinity of them since even slightly below T_f there is an infinity of valleys. It has been suggested by Mezard *et al* (1984a) that the valley division (merging) that occurs as T is lowered (raised) gives rise to a series of "micro phase transitions".

3.9 *Ultrametricity*

The hierarchy idea just mentioned has led to the introduction of a new concept into physics called *ultrametricity* (Mezard *et al* 1984a, b; for a review on the subject, see

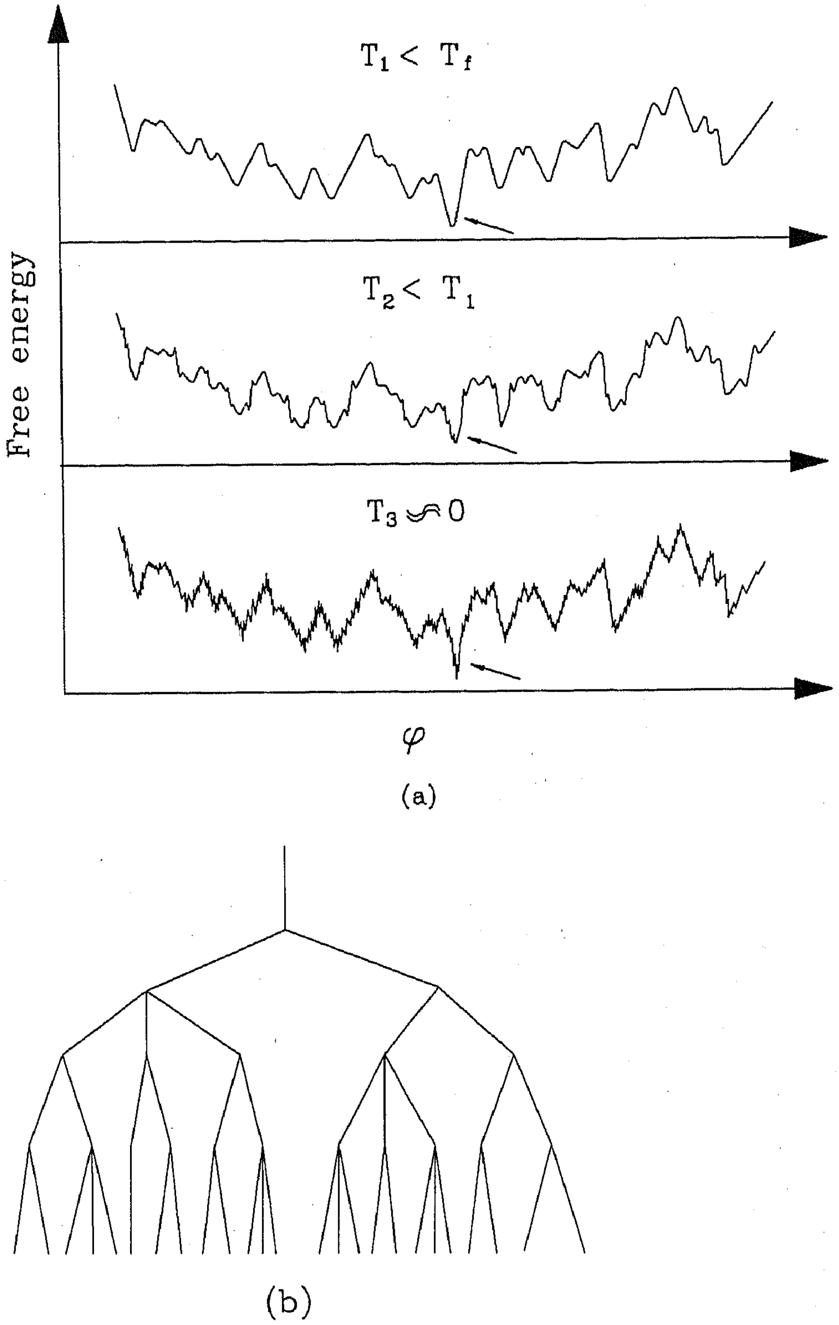


Figure 15. (a) schematic plot of the free energy hypersurface at various temperatures (recall also figure 1). Observe the development of "fractal structure" i.e., valley within valley etc. There is a notional minimum free energy state shown by the arrow. The barriers between the valleys are really infinite. (b) Tree showing valley bifurcation below T_f. Observe that the tree is inverted, with the "leaves" at the bottom.

Rammal *et al* 1986). The basic concept of ultrametricity belongs to mathematics, and interestingly it found an application in taxonomy long before being discovered by physicists.

Metric is a notion associated with distance. *A* metric space is a space endowed with a distance which in general obeys the triangle inequality i.e., if *A*, *B* and *C* are three points and if $d(A, B)$ etc., denote the distances between these points, then

$$d(A, C) \leqslant d(A, B) + d(B, C). \tag{44}$$

An ultrametric space is one where the inquality is stronger, being of the form

$$d(A, C) \leqslant \max\{d(A, B), d(B, C)\}. \tag{45}$$

Accustomed as we are to Euclidean spaces, an ultrametric space can cause severe problems in visualization. For example, if we consider a ball in an ultrametric space, then every point inside this ball is at the centre of the ball and the diameter of the ball is equal to its radius! It is comforting, however, that problems where the ultrametricity concept is actually utilized do not require such arduous visualization exercises, as the example to be presented now will illustrate.

Figure 16 shows a schematic evolution tree at the bottom of which are various species S1, S2··· etc. The branching points of the evolution are ordered and dated, and the vertical axis becomes equivalent to a time axis. A natural definition of distance between two species would be a quantity proportional to the age of the closest

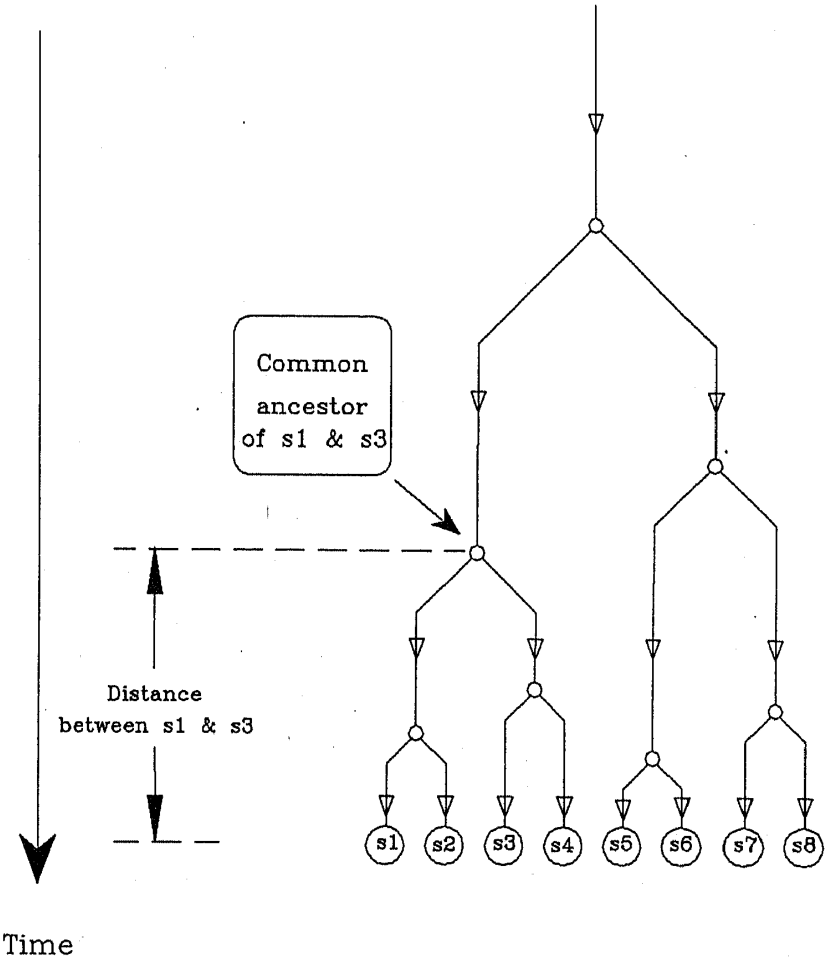
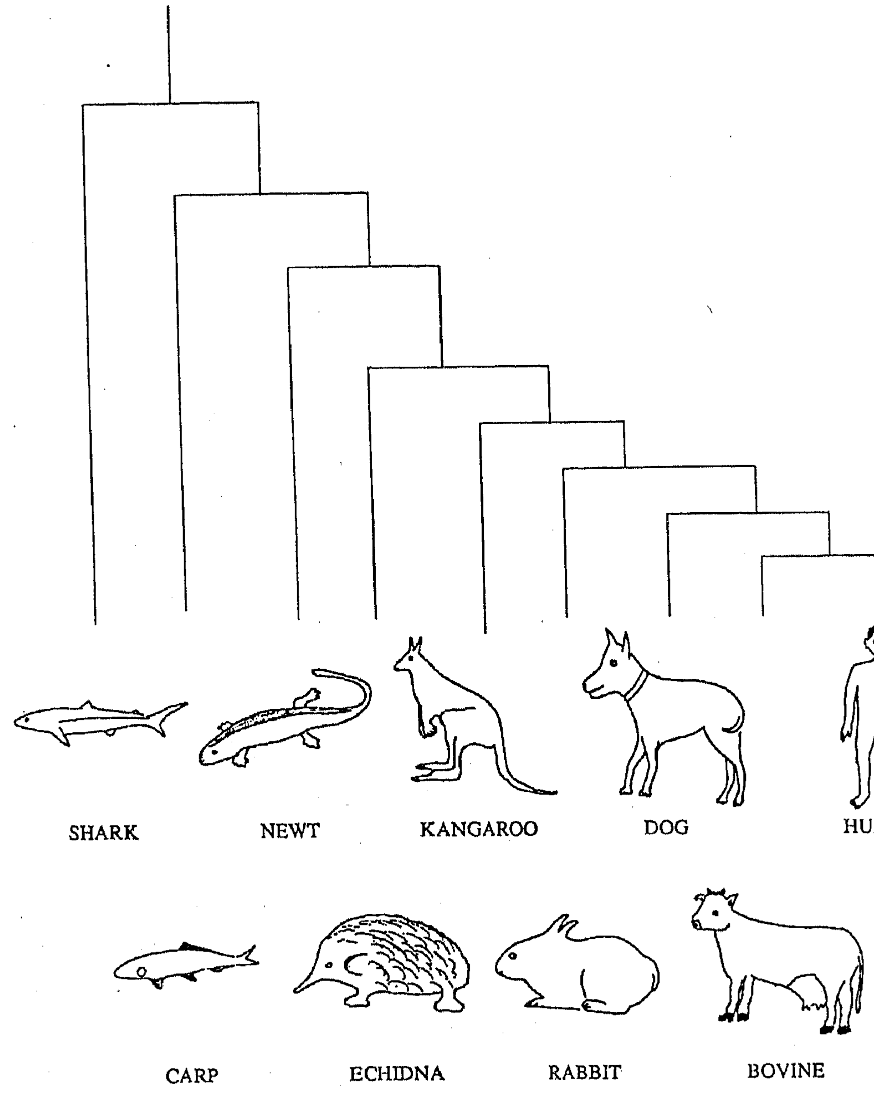


Figure 16. Schematic evolution tree showing the species $S_1, S_2 \ldots\ldots S_8$ which have evolved after a time T. The branch points at earlier times show the various ancestors. Distances between species S_i and S_j would be proportinal to the time that have elapsed since they branched off. An example is shown.

ancestor, that is, the time elapsed since they branched off. Distances measured in this fashion frequently obey the inequality (45), whence the 'space of species' may be regarded as being ultrametric.

A simple, tree-construction procedure devised by taxonomists is illustrated in figure 17(a), (b), the subject being genetic transformations in vertebrates as gleaned from differences in the α-hemoglobin chains. Each entry in the 9 × 9 matrix in figure 17(b) shows the amino acid differences between the α-hemoglobins of the two vertebrates involved. Starting with the last row, the matrix element 98 is clearly the smallest. Accordingly, as a first step, the human and the bovine are aggregated into a cluster and the matrix is renormalized into a (8 × 8) one. Here the entries that do not involve either human or bovine remain the same as before but where the latter two are involved, one selects the smaller of the two values. For example, considering the dog-human and the dog-bovine combinations, the corresponding numbers are respectively 28 and 23. Selecting the smaller of the two, we now say that the combination between dog on the one hand and the (human/bovine) cluster on the



17(a)

	1	2	3	4	5	6	7	8	
1	▨	85	84	84	80	75	80	75	Shark
2	85	▨	74	75	71	71	67	65	Carp
3	84	74	▨	71	67	69	65	62	Newt
4	84	75	71	▨	49	49	42	37	Echidna
5	80	71	67	49	▨	37	33	26	Kangaroo
6	75	71	69	49	37	▨	28	25	Rabbit
7	80	67	65	42	33	28	▨	23	Dog
8	75	65	62	37	26	25	23	▨	(Bovine–Human)
	Sha.	Carp	Newt	Ech.	Kan.	Rab.	Dog	(Bo. Hu.)	

	1	2	3	4	5	6	7	8	9	
1	▨	85	84	84	80	75	80	75	79	Shark
2	85	▨	74	75	71	71	67	65	68	Carp
3	84	74	▨	71	67	69	65	64	62	Newt
4	84	75	71	▨	49	49	42	43	37	Echidna
5	80	71	67	49	▨	37	33	26	27	Kangaroo
6	75	71	69	49	37	▨	28	25	25	Rabbit
7	80	67	65	42	33	28	▨	28	23	Dog
8	75	65	64	43	26	25	28	▨	17	Bovine
9	79	68	62	37	27	25	23	17	▨	Human
	Sha.	Carp	Newt	Ech.	Kan.	Rab.	Dog	Bov.	Hum.	

17(b)

Figure 17. (a) This figure based on the work of Kimura (1985 and quoted by Rammal *et al* 1986) depicts a branching as revealed by amino acid differences between the hemoglobin chains. (b) The differences are tabulated in matrix form. By systematically decimating the matrix as described in text, the evolution tree of a (a) can be built up.

other has a value 23. Likewise, we select 26 as the value for the Kangaroo-(human/bovine) combination. After the (8 × 8) matrix has been built up in this fashion, one takes up the next stage of clustering by enlarging the (human/bovine) cluster into the (human/bovine/dog) cluster leading now to a (7 × 7) matrix. In this way, by a step-by-step decimation procedure, the cluster tree can be built up.

Distance is a *dissimilarity* index. In many problems, the relationship between objects is better expressed via overlaps which is a *similarity* index. Let α, β, γ be three objects and $q_{\alpha\beta}$, $q_{\beta\gamma}$, $q_{\alpha\gamma}$ their overlaps. In terms of these, one has the metric inequality

$$q_{\alpha\gamma} \geqslant q_{\alpha\beta} + q_{\beta\gamma} - 1. \tag{46}$$

Likewise, the analogue of (45) is the ultrametric inequality

$$q_{\alpha\gamma} \geqslant \min\{q_{\alpha\beta}, q_{\beta\gamma}\}. \tag{47}$$

In a sense, test for ultrametricity boils down to evaluating triangle statistics. In an ultrametric space, the triangle (of overlaps) will all be either equilateral or isosceles, with the third side being larger than the two equal ones. For systems with a large number of degrees of freedom, ultrametricity is a natural type of organization.

Ultrametricity ideas have been applied both to TAP states/valleys as well as to replicas (Mezard *et al* 1984a, b). Naturally the two descriptions are related. Consider three pure states s, t and u. Mezard *et al* show that the overlaps q_{st}, q_{tu} and q_{su} obey the rule (47), whence the space of TAP states has an ultrametric structure. Now suppose $q_{tu} = q_{su} \equiv q \leqslant q_{st}$. We can then put the states s and t into a cluster and place u outside of it. The idea is to have groupings such that states *within* a cluster have larger overlaps among themselves than with respect to a partner *outside* the cluster. Using this guideline one can build up a hierarchy of clusters with an underlying tree structure—see figure 18.

The replicas can also be given a similar hierarchical structure. It turns out that M_k the probability that k *states* are in the same cluster also describes the probability of having k different *replicas* within the same cluster. Naturally this probability would depend on the overlap scale q chosen. This opens up the possibility of giving a detailed, quantitative interpretation to the idea of micro phase transitions introduced earlier. We now observe that as the temperature is varied, the system organizes its states to achieve overlaps on a progressively varying q scale.

This concludes our survey of the spin glass problem, and we are now ready to turn to other related topics.

4. The travelling salesman problem

4.1 *The statement of the problem*

The travelling salesman problem (TSP) may be stated as follows: A list of N cities is given, together with the distances between all pairs of them. A salesman is required to perform a tour by starting from any city and returning to the original place of departure, visiting all the remaining cities *en route* once and only once. What route should he choose in order to minimize the total distance travelled? In slightly more formal terms, one is asking for the shortest closed path through a given set of N

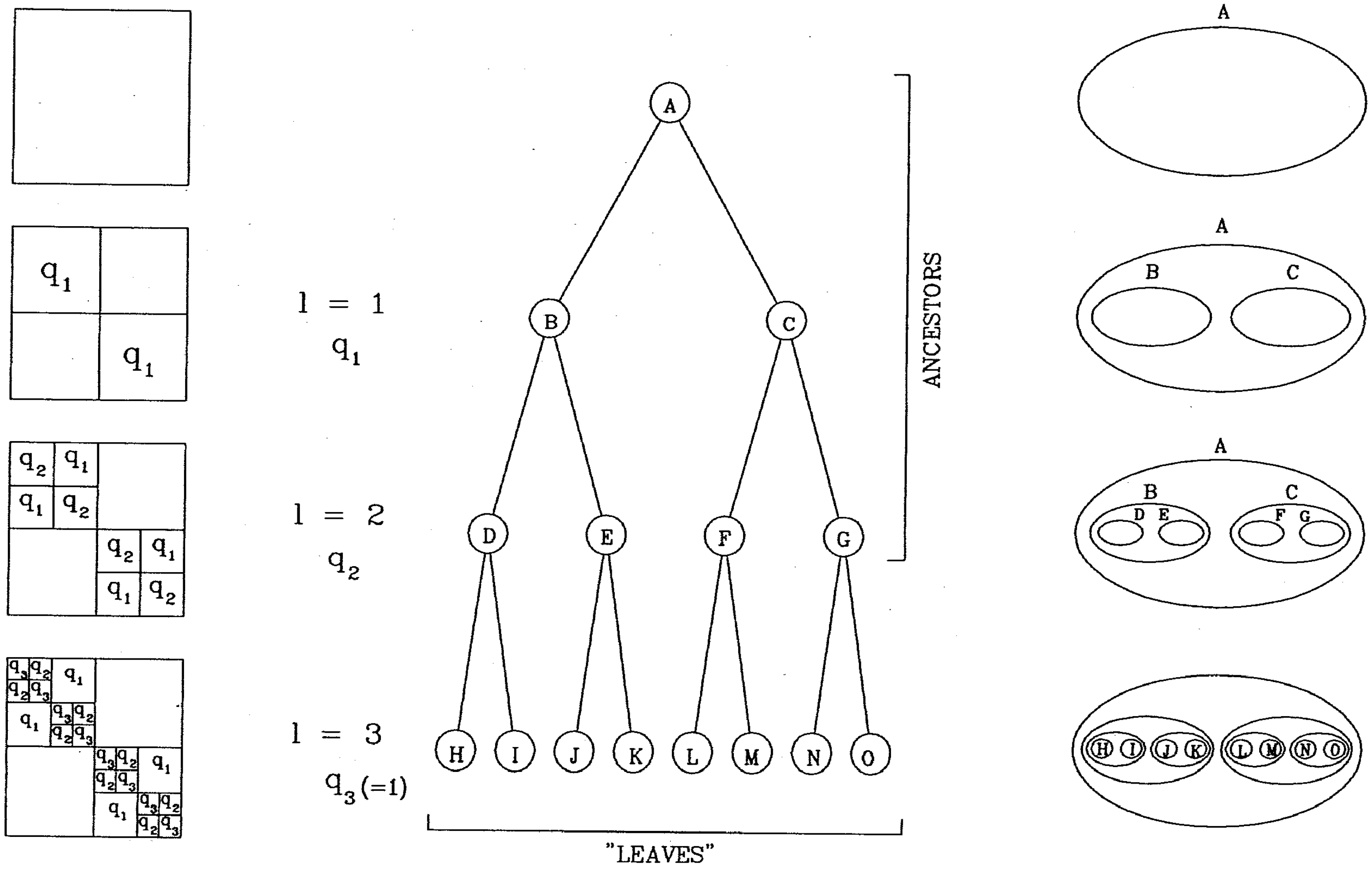


Figure 18. This figure illustrates schematically and with an elementary example, the relationships between the matrix partitioning scheme of Parisi, the branching and the clustering of states. The states of the system are the "leaves" at the bottom i.e. at the level $l = 3$. These are clustered as shown alongside at the right. The clustering is according to the overlaps which decrease as the meeting between leaves occurs at distant ancestors.

points, and in this form the problem occurs frequently in several contexts. To put it in slightly different terms, this is an example of a *combinatorial* optimization problem.

Mathematically, the TSP can be formulated as follows: Let l_{ij} denote the distance between cities i and j and L be the length of the tour i.e., $L = \Sigma_i l_{i,i+1}$, with $(N, N+1) = (N, 1)$. Further, let P denote a permutation of the original order of cities, with $P(i)$ denoting the new position of a city which occupied position i in the original list. With these definitions, the TSP implies finding the permutation P for which

$$L_P \equiv \sum_{i=1}^{N} l_{P(i),P(i+1)}, \quad P(N+1) = P(1) \tag{48}$$

is a minimum. The quantity L is sometimes referred to as the *cost function*, although in real life the cost of travel between cities i and $i+1$ may depend on factors other than distance e.g. mode of transport. Such complications can and do occur but the problem stated above is the 'hydrogen atom' version, and hence the wide interest in it.

The cost of a tour is the analogue of the energy of a particular configuration in statistical mechanics, while the counterpart of the phase space is the space of all tours. In the same language, the optimal tour corresponds to the ground state.

4.2 *On exponential-time algorithm*

The TSP is a paradigm beloved of workers in Operations Research, providing as it does a useful test bed for checking out optimization algorithms. This is a convenient juncture to make a few qualitative remarks on 'intractability' and related matters (Garey and Johnson 1979). For our purposes, a problem may be regarded as a general, many-parameter question which is to be answered. A specific set of values assigned to these parameters defines a given *instance* of the problem. We remind ourselves that algorithms are step-by-step procedures for solving problems and, in the context of this paper, simply mean computer programs. An algorithm is said to *solve* a problem if it can provide solutions for *every* instance one can associate with the problem. For example, in the N-city TSP there are $N(N-1)/2$ inter-city distances, and a successful TSP algorithm must be able to find the optimal solution (i.e. minimal length tour) for *all* configurations of the cities. Naturally the time to find this solution would increase with the "size" n of the problem, a convenient measure of which for the TSP would be $(N-1)!/2$. The *time complexity function* for an algorithm expressed as a function of the size n, is the largest amount of time required by the algorithm to solve the problem.

It is customary to distinguish two cases (i) where the time required for the execution of the algorithm varies as n, n^2, $n^3, \ldots$ etc., and (ii) where the time variation is like 2^n or 3^n or $n!$ etc. The former are called polynomial-time algorithms and the latter as exponential-time algorithms. The implications of such a classification may be better appreciated by referring to table 2.

A problem is deemed to be well solved only if a polynomial-time algorithm is known for it. It is essential that this algorithm works for *every conceivable instance of the problem.* It sometimes happens that while the polynomial-type is adequate for most instances, there are nasty exceptions which require exponential-time algorithms. In general, a problem is deemed intractable if no polynomial-time algorithm can possibly solve it. A polynomial-time (P) problem is one which can be solved by a

Table 2. Size of the problem corresponding to different levels of complexity which can be handled by improved computers. The problem size is computed from the relation $(N_1/S_1) = (N_2/S_2)$ where N_1 and N_2 denote the *total* number of operations performed by two computers of speeds S_1 and S_2. In all cases, it is assumed that the size handled by the best *present* computers is n.

Complexity	Largest size with a computer 100 times faster	Largest size with a computer 1000 times faster
n	$100\,n$	$1000\,n$
n^2	$10\,n$	$31{\cdot}6\,n$
n^3	$4{\cdot}64\,n$	$10\,n$
n^5	$2{\cdot}5\,n$	$3{\cdot}98\,n$
2^n	$n+6{\cdot}64$	$n+9{\cdot}97$
3^n	$n+4{\cdot}19$	$n+6{\cdot}27$
$f(n)$	$f^{-1}[100\,f(n)]$	$f^{-1}[1000\,f(n)]$

polynomial algorithm. A NP-complete problem, on the other hand, is one which cannot be solved in polynomial time. While a more rigorous definition of NP-completeness is certainly possible, the above is adequate for our present purposes. Viewed in this light, the TSP is a NP-complete problem.

4.3 *Algorithms for tackling the TSP*

Although the TSP is a NP-complete problem, many attempts at algorithmic solution have nevertheless been made. Before describing these, reference is necessary to a bound given a long time ago by Beardwood *et al* (1959). They showed that the asymptotic value of the shortest tour length is given by

$$\lim_{N\to\infty} L_{\min}/N^{1/2} = \gamma\sqrt{2}\int_D [\mu(x_1, x_2)]^{1/2}\,\mathrm{d}x_1\,\mathrm{d}x_2 \tag{49}$$

where $\mu(x_1, x_2)$ describes the distribution of the N points in the bounded region D, the distribution of points being assumed to be random. The constant $\gamma \sim 0{\cdot}53$.

Turning to the algorithm, we start with one described as *greedy*. This algorithm does not in any sense attempt an optimal solution but its virtue is that it is fast. To implement it, one starts from a city picked at random, moves to the city nearest to the starting one, makes the next step by going to the closest remaining city, and so on. The program results in N tours, usually not all distinct. Typically the greedy algorithm yields tours several times costlier than the optimal one. It is thus useful for obtaining an order-of-magnitude estimate.

Iterative search is a better strategy. In a sense, it is a heuristic approach. Before describing it, the concept of a *λ-optimal tour* must first be introduced. A tour is said to be λ-optimal (or λ-opt), if it is impossible to obtain a tour with smaller cost by replacing any set of λ links by any other set of λ links. (A link denotes the straight line path between two specified cities.)

Consider now 2-opt tours. As just noted, the aim is to start from a given tour T

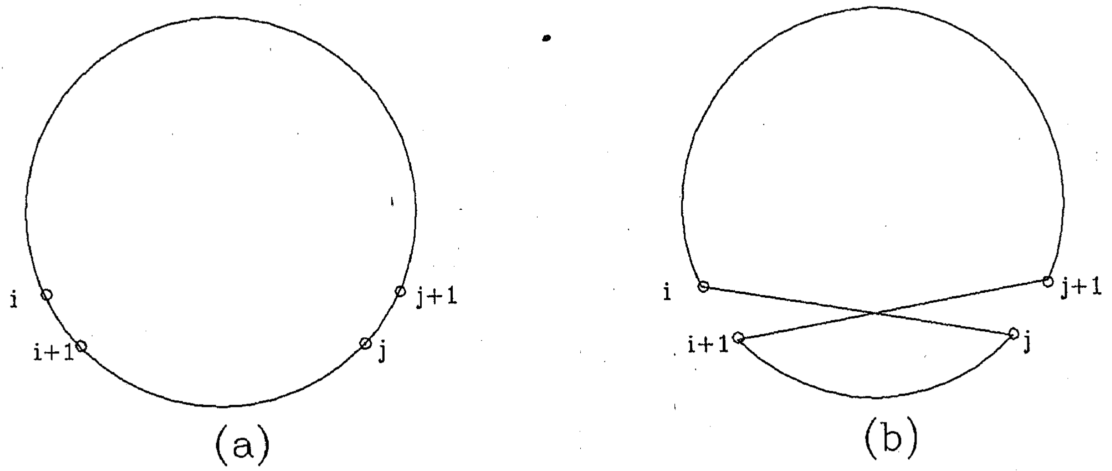


Figure 19. Illustration of the simplest move set. It involves two links.

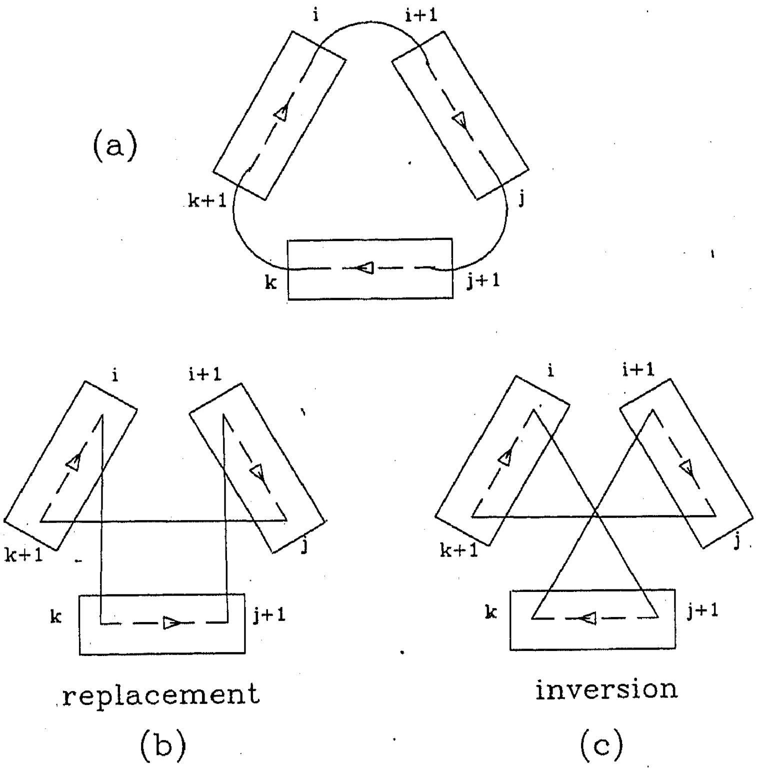


Figure 20. Some examples of 3-move set.

say, and then construct another tour T' slightly different from T and less costly. The simplest move set for going from T to T' involves 2 links, the change in the linkages being effected as illustrated in figure 19 (Lin 1965). The change is sometimes referred to *inversion* since part of the tour (i.e., that between $i+1$ and j) is performed in the reverse direction. Next in complexity is the 3-move set, two versions of which are illustrated in figure 20. One of these is a *replacement* while the other is a *reversal* (Lin 1965). Further refinements have been suggested by Lin and Kernigan (1973). The Lin algorithm is essentially iterative, following a tour sequence T, T', T'', $\cdots$ steadily decreasing in cost till no further decrease occurs.

It should not be assumed from the above that such a strategy automatically leads to the global minimum. It does not; instead, it usually leads to a local minimum. In the early eighties, several workers (Bonomi and Lutton 1984; Cerny 1985; Kirkpatrick 1984) almost simultaneously recognized that iterative improvement is rather like rapid quenching, whence a trapping in a local valley is almost inevitable. A natural cure is to employ annealing (in conjuction with move sets as in the Lin algorithm), which is precisely what the Metropolis algorithm facilitates. And thus emerged the *simulated annealing method* which has become quite popular in recent years. This naturally required the association of a temperature with the problem. However, in this case temperature lacks physical significance, being merely a convenient parameter. It will of course be noticed that simulated annealing implemented with T set equal to zero reverts back to the iterative algorithm.

One would naturally like to know how effective simulated annealing technique is in problems that are NP-complete. Some indications of this are available from numerical studies on the spin-glass problem. The lowest energy state obtained in large samples is $\sim -0{\cdot}7$, as compared to the ground state energy $U(0) = -0{\cdot}763$ obtained by analytical studies invoking replica-symmetry breaking. In short, while simulated annealing does produce a low-lying state, one does not know whether the result is the optimal solution or how far (in general) it is from this solution. However, in large problems with built-in frustration, one expects the cost landscape to be rather like that in the spin-glass problem whence, to leading order in N, the solution obtained via simulated annealing can be approximated to the global minimum.

4.4 *Some results obtained using simulated annealing*

A simple illustration of the way simulated annealing operates is given in figure 21 which presents results similar to those first reported by Cerny (1985). In this case, the salesman has to tour 100 cities distributed uniformly on the unit circle. The optimal tour is trivial, namely the path going round the circle. Nevertheless, to show how simulated annealing works, we start with a tour as in figure 21(a) and an initial temperature of 1·0°. Temperature is decreased linearly with time, and tours after every 1000 Monte Carlo steps are plotted. We observe that the simulated annealing algorithm is able to hunt the optimal solution with ~ 16000 permutations as against a total of $99!/2$. One cannot naturally overplay this example since the greedy algorithm will yield the optimum tour in a hundred steps; rather, our aim is to illustrate how annealing produces changes.

Perhaps a better example is furnished by the results shown in figure 22, the like of which was first reported by Kirkpatrick (1984). In this figure, we have 100 cities arranged in a square lattice. One starts with a tour as in (a), an initial temperature of 1, and employs an exponential cool down. The sequence of pictures in the figure are very reminiscent of the gas → liquid → solid transition.

In real life, cities are not arranged on a lattice but dispersed somewhat randomly. Results for random city TSP are shown in figure 23. Once again we have 100 cities, and an exponential cool down. The low-temperature tour has a respectable appearance with no bond intersections. By contrast, the high-temperature tour resembles particle motions in a gas.

Researchers often study the TSP with large values of $N(\sim 10{,}000)$ in order to expose the complexities underlying the problem. Of course no salesman ever makes a round

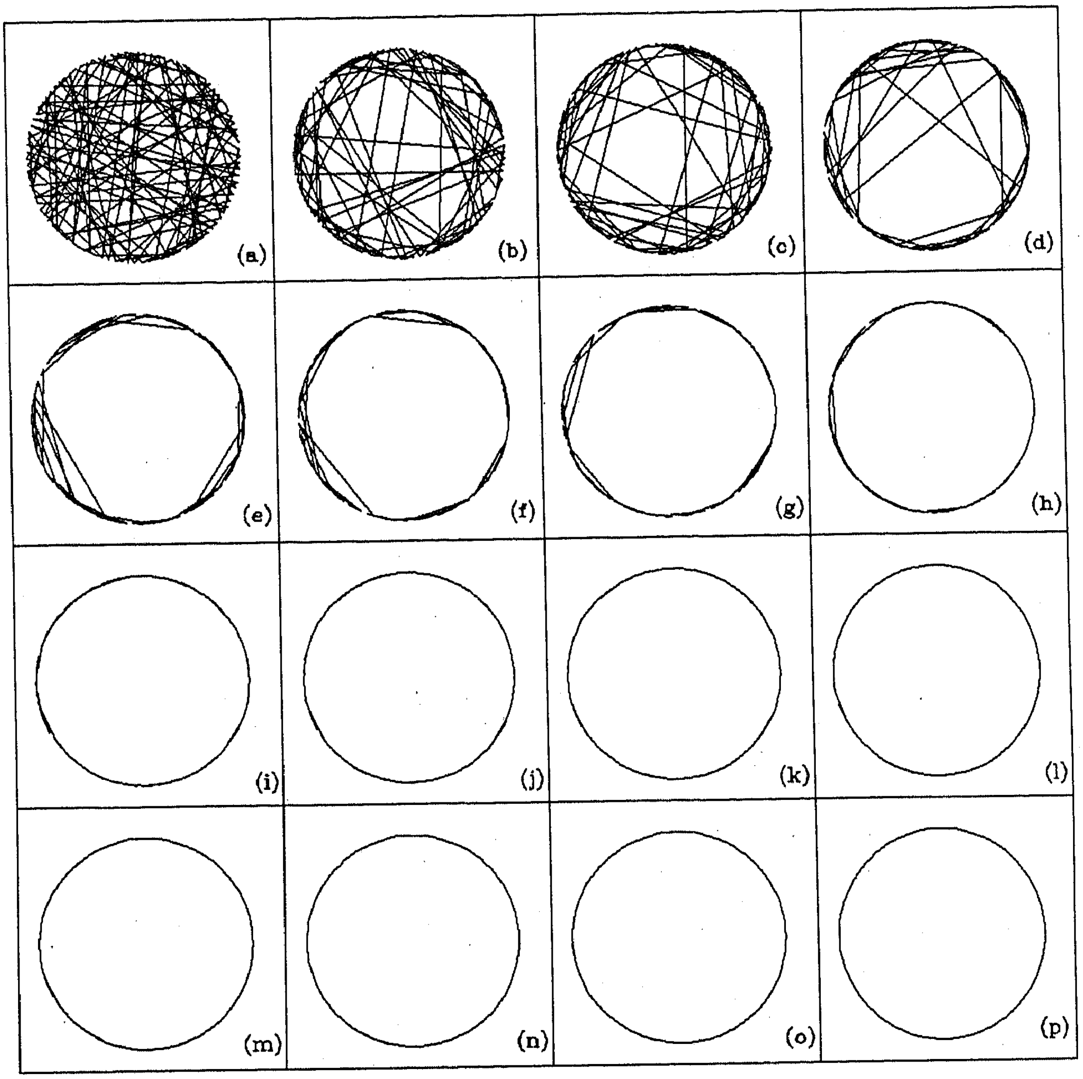


Figure 21. Optimization of a 100-city circular tour by simulated annealing. The starting tour is shown in (a). Subsequent frames show the tour after every 1000 Monte Carlo steps.

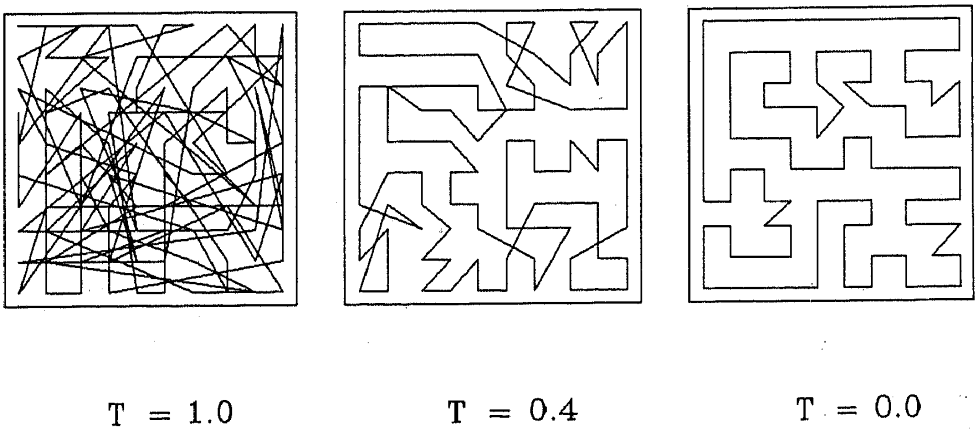


Figure 22. TSP on a 10 × 10 grid, as studied by simulated annealing.

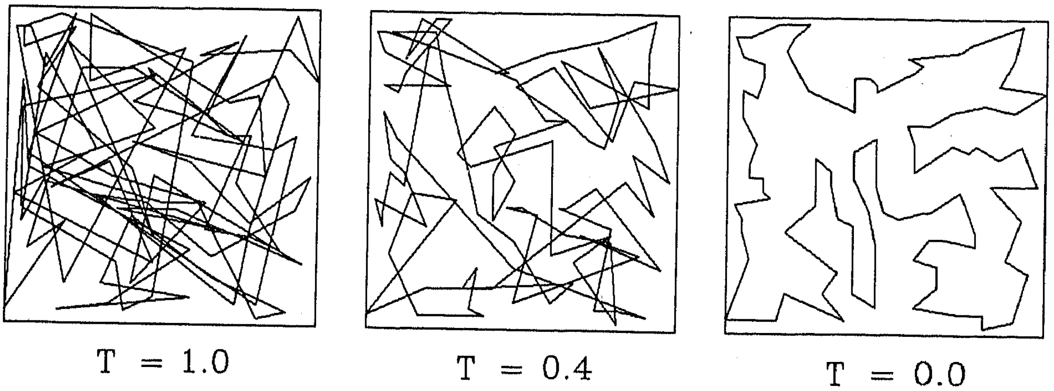


Figure 23. Random city tours as studied by simulated annealing.

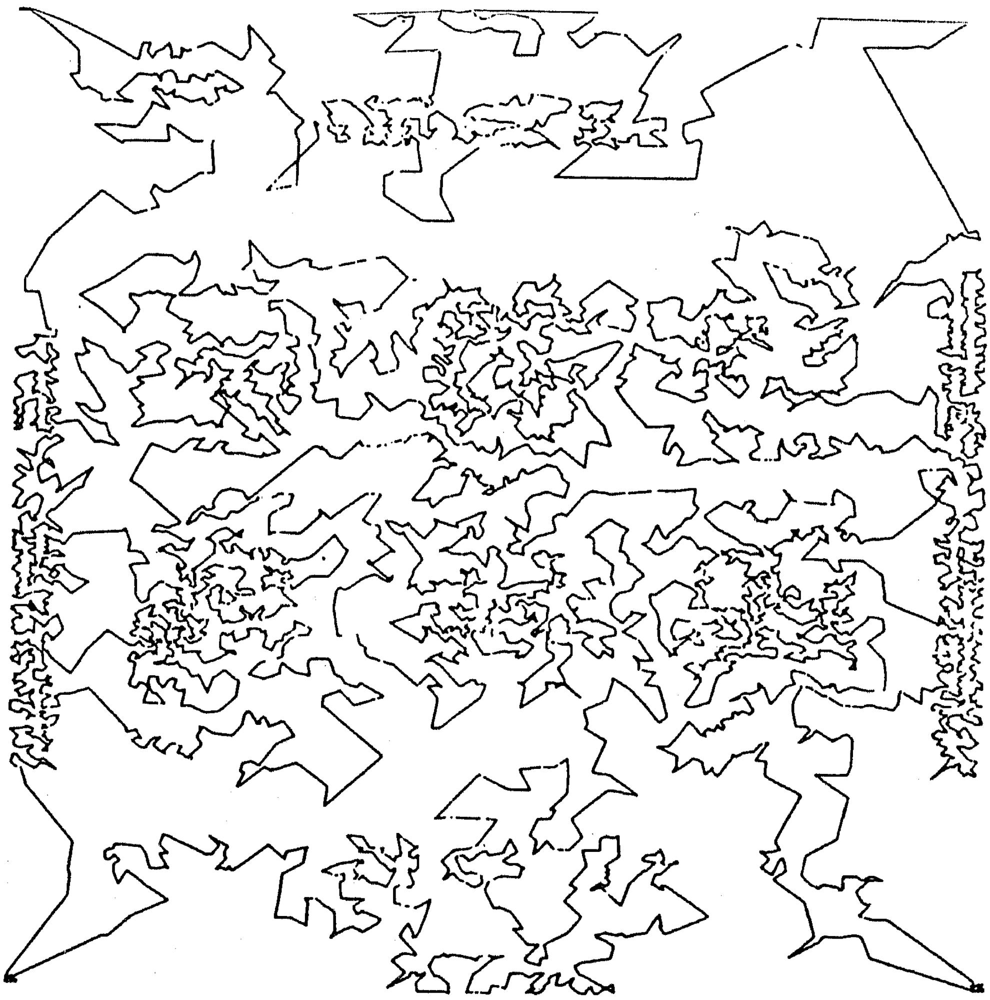


Figure 24. A TSP solution for drilling 6406 holes in a PCB (after Kirkpatrick 1984).

trip involving thousands of cities! And yet, the analysis of such problems does have practical relevance, though not with respect to sales tours. One example is the automated drilling of holes in a printed circuit board (PCB), where the movement of the drilling element plays the role of tours performed by the salesman. A difficulty with this application is that the time taken to drill a hole usually exceeds the time of transit between drilling positions, whence it is not worthwhile to minimize the 'tour' duration. However, with the advent of laser drilling, the time to drill a hole is practically negligible, and the application of combinatorial optimization to PCB drilling begins to make sense. Figure 24 shows an example of such a result obtained by Kirkpatrick (1984). In this case there were 6406 holes to be drilled, and the "tour" was obtained by first applying the greedy algorithm and then simulated annealing.

In general, simulated annealing is useful in complex optimization problems characterized by complicated energy (i.e., cost landscape, the main feature of which is the presence of a large number of local valleys (metastable states)). The algorithm is reasonably efficient and adequate if the cost associated with the lowest valley (ground state) is not very much smaller than those of other minima (metastable states). In problems which have a phase transition, the simulated annealing technique must be used with care since near the transition temperature T_c, there would be slowing down effects analogous to critical slowing down in the phase transition problems of condensed matter physics.

4.5 *Statistical mechanics of the TSP*

The application of the simulated annealing technique to TSP has encouraged a study of the statistical mechanics aspects also. Considering that the introduction of a temperature has itself the appearance of being artificial, one might wonder what statistical mechanics means in this case. Hopefully the answer to this would become clear shortly when we consider the details.

We begin our discussion of the statistical mechanical aspects by noting that the TSP has elements of frustration, associated with a conflict "between the short-range requirement that each step of the path be as short as possible and the long-range requirement that every point be visited once and the path be closed" (Kirkpatrick 1981). One would like to know the consequences of this frustration, and this is accessible only via a statistical mechanical type of discussion. The latter will necessarily bring in concepts like the partition function, the free energy etc., and it is natural then to expect a free energy landscape as in figure 1(c), the horizontal axis denoting the space of all possible tours (associated with a given instance). The valleys are the analogue of TAP states for this problem, and, depending on the initial conditions, one would be trapped in one or the other of these valleys. Incidentally, since these are more or less alike, it is not worth hunting for the lowest, although one among them may be notionally so. In other words, the worst case is not all that significant, which is comforting!

We turn now to some of the studies which have been carried out. As formulated in (48), the TSP is a *random-point problem* in that the positions of the cities are chosen at random in a plane (in general in a D-dimensional space). There is also another version known as the *random-link problem* where the inter-city length $l_{ij}(=l_{ij})$ is treated as a random variable having a specified distribution $\rho(l)$. The latter model has been studied by Kirkpatrick and Toulouse (1985) and by Vannimenus and Mezard (1984). The former authors assumed $\rho(l)$ to be constant in the (normalized) interval $0-1$, whereas the latter assumed.

$$\rho_r(l) = (l^r e^{-k}/r!). \tag{50}$$

Some results of this model map on to the D-dimensional random point model with the identification $r \rightarrow D-1$. The study of Vannimenus and Mezard showed that there are two regimes of temperature, a high-temperature ($\mathscr{H}$) one and a low-temperature ($\mathscr{L}$) one. In the $\mathscr{H}$-regime, the annealed approximation is made to write the partition function as

$$[Z]_{\mathrm{Av}} = \int \prod_{i<j} \rho(l_{ij})\,\mathrm{d}l_{ij} \sum_{P} \exp\left(-\beta \sum_{i} l_{P(i),P(i+1)} \right) \tag{51}$$

The integral over l_{ij} gives the same result for every permutation P so that

$$[Z]_{\mathrm{Av}} = N!\left(\int_0^\infty \mathrm{d}l\rho(l)\exp(-\beta l) \right)^N = N!(g(\beta))^N, \tag{52}$$

where $g(\beta)$ is the characteristic function of $\rho(l)$. This then yields the annealed free energy

$$F_{\mathrm{ann}} = -\ TN\ln(N/e) - \mathrm{TN}\ln g(\beta) + O(\ln N), \tag{53}$$

and the annealed average length

$$L_{\mathrm{ann}} = -N\mathrm{d}\ln g(\beta)/\mathrm{d}\beta. \tag{54}$$

The assumption underlying (51) implies certain scaling behaviour for the average length $\langle L \rangle$, the entropy and the energy, when $N \rightarrow \infty$ (T fixed). For example, in the $\mathscr{H}$ – regime, $\langle L \rangle$ varies as

$$\langle L \rangle \sim \mathrm{Na}$$

where a is the edge of the D-dimensional cube. This implies that the average bond length is a and this is what (54) predicts. However, in the $\mathscr{L}$-regime, the average bond length will clearly be not that large. Rather one expects it to be $(V/N)^{1/D}$ where $V = a^D$ is the volume of the cube. Thus

$$\langle L \rangle \sim aN^{1-(1/D)}$$

which shows that the annealed approximation will not do.

The signals are clear. Any attempt to deduce analytical results must bear in mind the fact that (i) optimization implies search for solution in a $T=0$ situation, and (ii) on account of inherent frustration (already drawn attention to), the cost landscape will have characteristic rugged features. The analytical methods employed must therefore be tuned to these aspects. Thus, Vannimenus and Mezard observe that to obtain meaningful results one must average the logarithm of the partition function over all instances i.e., one must perform configurational average. This, however, is a formidable task, and they therefore content themselves with placing some bounds on L_{min}.

A very interesting model results when $r \rightarrow \infty$ in (50). One then obtains the infinite-dimensional random – point model which is soluble, and predicts a phase transition.

Going back to the finite r case, why not use the replica trick? This is precisely what Orland (1985) did but he was unable to find a (stable) solution. He conjectures that replica symmetry breaking might be required. No clear phase transition is

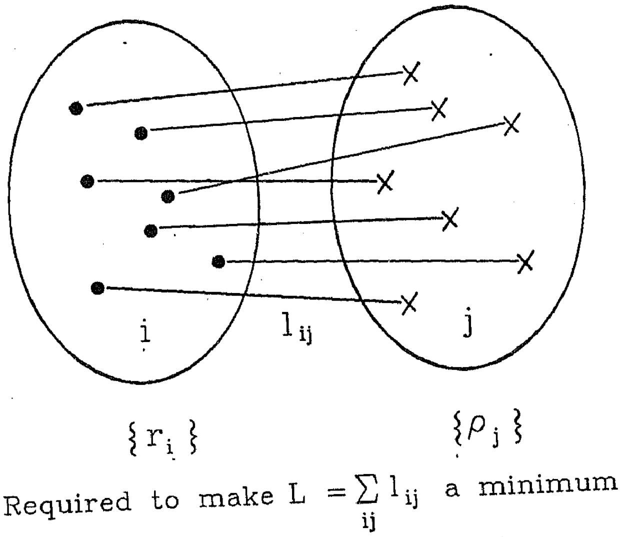


Figure 25. In the BMP, one has two sets of points and they are paired off. The problem is to find the pairing scheme for which the sum of all bond lengths is a minimum.

predicted (unlike in the $r \to \infty$ case) but there are two regimes $\mathscr{H}$ and $\mathscr{L}$, and a cross over between them.

A variant of the TSP is the Bipartite Problem (BMP). In this, one has two (equal) sets of points $\{r_i\}$ and $\{\rho_i\}$ in a D-dimensional space, and the problem is to find the best matching between the sets i.e., one must construct bonds as in figure 25 such that the *total* bond length is a minimum. As in the TSP, one can make either the points $\{r_i\}$ and $\{\rho_i\}$ random or the inter-set linkages l_{ij} random. In the latter case of course, one has to specify the distribution $\rho(l_{ij})$. Either way, the matrix l_{ij} defines an instance of the problem.

Orland (1985) observes that unlike the TSP, the random-point BMP belongs to the P-class. He speculates that P-class problems would not show replica symmetry breaking whereas the NP-class problems (e.g. the TSP) would.

Mezard and Parisi (1985) have investigated the random-link version of the BMP. The order parameter is now a tensor, much more complicated than in the spin glass case. Assuming $\rho(l)$ to be as in (50), Mezard and parisi attempt a replica-symmetric solution for finite r. They do not observe a phase transition but a freezing nevertheless.

The case $r = 0$ is interesting, corresponding as it does to a non-vanishing probability of having infinitely short links. Mezard and Parisi assume that replica symmetry is not broken but nevertheless find that the solution turns out to be a whole function $G_r(l)$ (rather like Parisi's $q(x)$). This immediately raises the question whether there is ergodicity breaking as in spin glass and, to find the answer, Mezard and Parisi studied the overlap probability function $\mathscr{P}_{\{l_{ij}\}}(q)$, where q now refers to the overlap between two copies of the system corresponding to the same instance $\{l_{ij}\}$ but with different bipartite links. They remark that a nontrivial $\mathscr{P}_{\{l_{ij}\}}(q)$ function is possible only with replica symmetry breaking. Accurate numerical work which can decide whether $\mathscr{P}_{\{l_{ij}\}}(q)$ is indeed so, is lacking at present. Meanwhile, the existence of the function $G_r(l)$ strongly suggests that replica symmetry *is* broken.

Extensive numerical work related to the statistical mechanics of the TSP has been carried out by Kirkpatrick and Toulouse (1985) who have studied various features like overlap, ultrametricity etc. Let us start with overlap. Consider two tours α, β

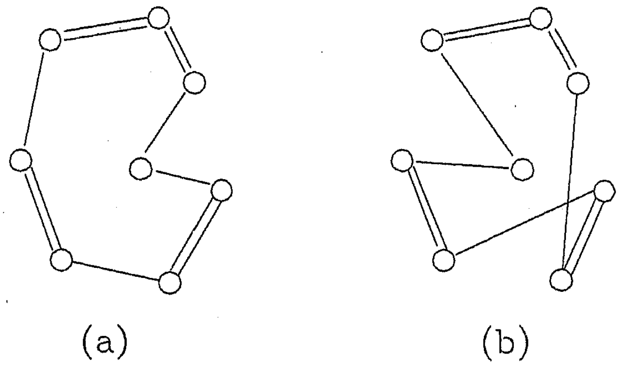


Figure 26. (a) and (b) show two 8-city tours α and β say. The common bonds are shown by double lines. Overlap $q_{\alpha\beta}$ is computed from the fraction of bonds common to the tours.

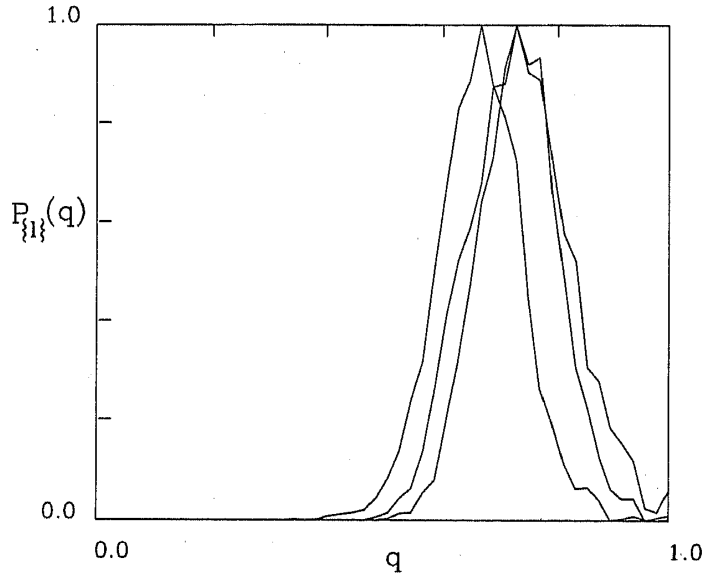


Figure 27. Overlap probability distribution $\mathscr{P}_{\{1\}}(q)$ for a random-link model 50-city tour. Results are shown for three different instances of the problem. In each case, hundred 2-optimal tours were generated.

corresponding to the same instance $\{l_{ij}\}$. The overlap $q_{\alpha\beta}$ is defined to be the fraction of bonds common to both tours, without regard to the direction in which they are traversed (see figure 26). To construct $\mathscr{P}_{\{l_{ij}\}}(q)$ for a N-city problem, one first performs several simulated annealing runs corresponding to different initial conditions. If M tours have been generated, then ${}^{M}C_2$ tour pairs can be formed which can then be analyzed for overlaps. Typical histograms are shown in figure 27.

Another quantity of interest is the bond frequency distribution $p(f)$. In a given instance of a N-city tour, $N(N-1)/2$ inter-city links or bonds are possible. In a sample of M tours say, not all these bonds will make their appearance. Suppose now that bond i occurs N_i times. The relative frequency f_i for its occurrence is then (N_i/M). The distribution $p(f)$ is defined as

$$p(f)=\sum_i \delta(f-f_i). \tag{55}$$

A typical result from a numerical study is shown in figure 28.

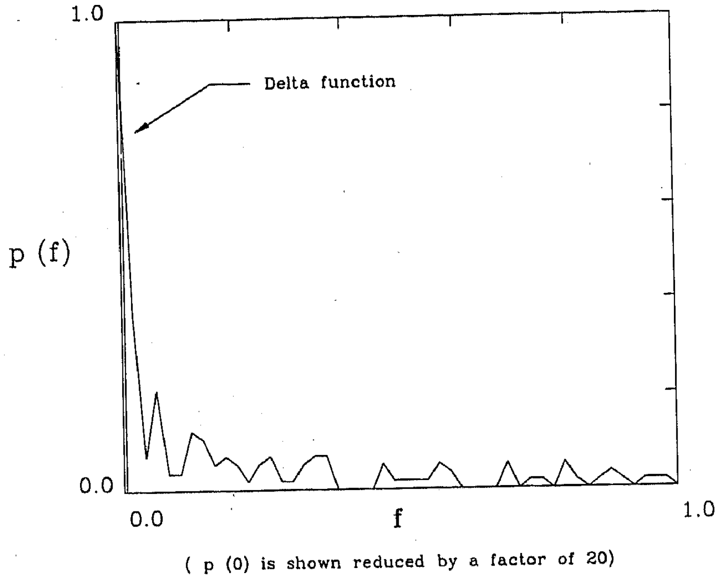


Figure 28. Bond frequency distribution $p(f)$ for one instance of a random-link model 50-city tours. Computation is carried over a set of hundred 2-optimal tours.

The moments of $p(f)$ are related to overlaps. For example,

$$q = \frac{1}{N}\sum_i f_i^2 \to \frac{1}{N}\int_0^1 \mathrm{d}f\, f^2 p(f). \tag{56}$$

Similar expressions may be written for higher-order overlaps involving, $3, 4 \cdots$ etc., tours.

Guided by numerical results, Kirkpatrick and Toulouse (1985) suggest the *ansatz*

$$p(f) = Af^{-\gamma} + \{N(N-1)/2 - A/(1-\gamma)\}\delta(f) \tag{57}$$

where $\gamma < 1$ and A are suitable constants. They remark that this form for $p(f)$ implies that only a finite number of bonds per city participate in the locally stable solutions of a TSP, a hypothesis which suggests intriguing heuristics for restricting the search for an optimal solution of the TSP.

Tests for ultrametricity have also been performed. Basically one generates M tours corresponding to a given instance, and then computes overlaps $q_{\alpha\beta}$ for all (α, β) combinations (of course, $\alpha \neq \beta$). If the space of overlaps is ultrametric, then the triangles formed by $q_{\alpha\beta}$, $q_{\beta\gamma}$ and $q_{\alpha\gamma}$ will be either equilateral or isosceles (with the third side being larger than the other two). Representing the shorter overlaps along the x- and the y-axes, one obtains distributions as in figure 29 from numerical studies. Certainly there are strong suggestions of ultrametricity.

One can now take stock and summarize as follows: For finding answers to *specific instances* of the TSP, we must continue to rely on available algorithms. The injection of annealing into such algorithms has undoubtedly improved matters compared to what obtained before. Thus one can not only escape local minima but also converge to a "reasonably good" solution fairly fast. Of course getting at the truly *optimal* solution is a different matter, and it is safe to assume it will elude detection since the problem belongs to the NP class. On the other hand, given the existence of elements of

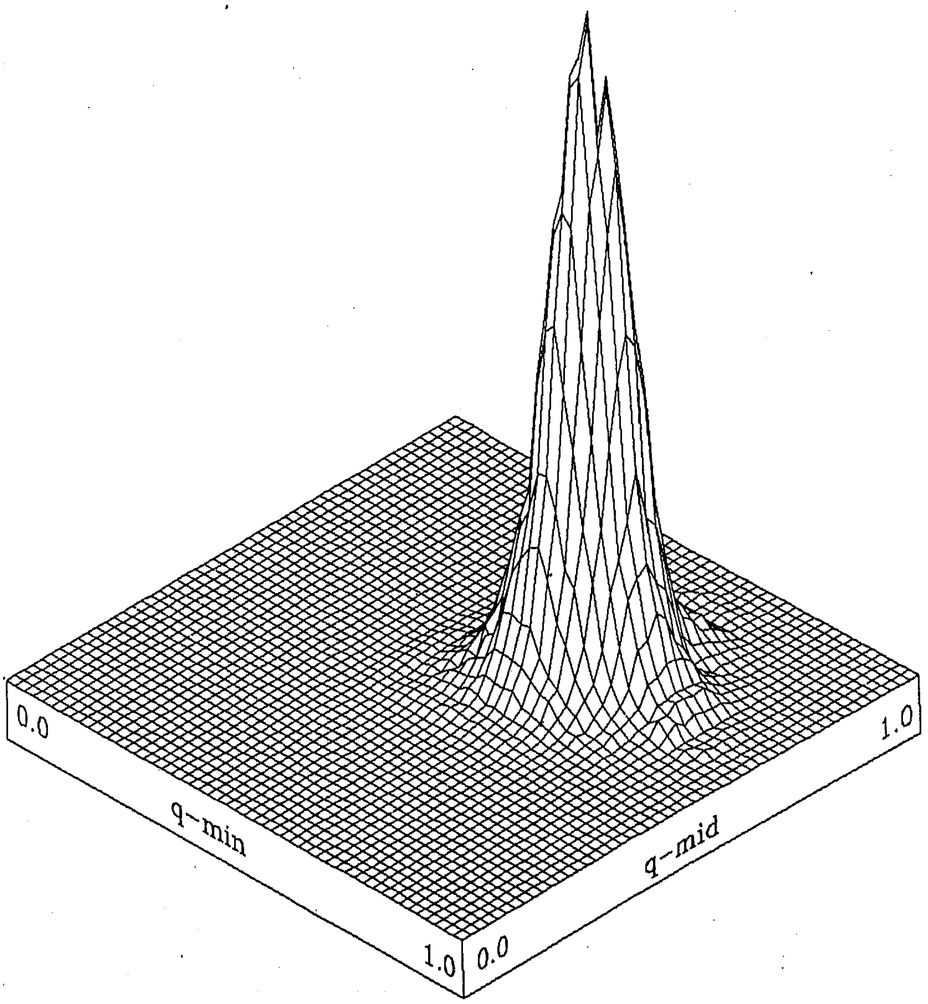


Figure 29. Triangle statistics for hundred 2-optimal 50-city tours. As explained in the text, overlaps $q_{\alpha\beta}$ are first computed. Next one considers the middle and the smallest sides of overlap triangles, and plots their distribution. Perfect ultrametricity would yield a 2-*D* peak along a 45° line. Our results show a hump, indicative of partial ultrametricity.

frustration in the TSP, a solution degeneracy is natural, whence it is meaningful to claim that a 'reasonably good' solution obtained as described above will be only marginally higher in cost compared to the truly optimal one, and therefore acceptable. Thus, at the practical level, the worst case becomes an academic issue.

As far as statistical mechanics is concerned, it enables one to look at the *family* of solutions as a whole and get a feel for how the configuration landscape changes as the parameters defining an instance are varied. For instance, there are the fluctuations in $P_{\{l_{ij}\}}(q)$ which give some measure of these changes. Kirkpatrick and Toulouse (1985) refer to this as the "plasticity" of the landscape, a knowledge of which would permit "gardening"! Indeed, gardening is even more important in the design of memories, as we shall find in § 7.

4.6 *Feedback to spin glass*

In the examples cited thus far, there has been a flow of ideas from the spin glass field to others. We now discuss a case of reverse flow in which an optimization strategy

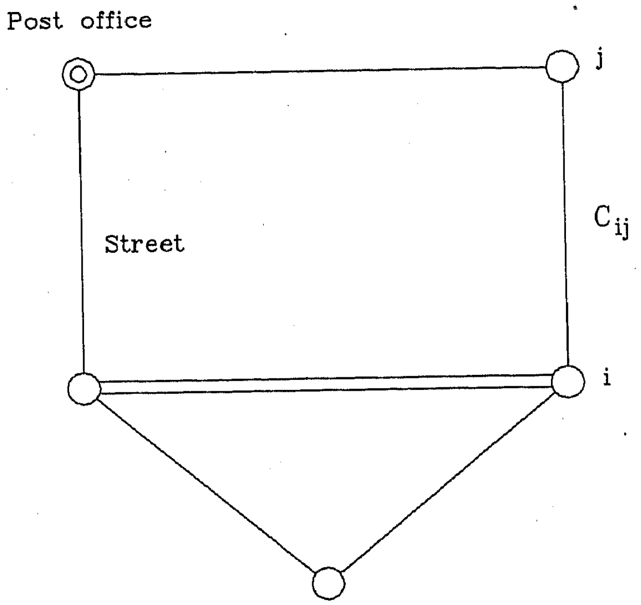


Figure 30. The Chinese postman's problem. For explanations, see text.

derived from graph theory has been exploited to obtain a knowledge of the ground state of spin glass. Two studies of this nature have been reported (Bieche *et al* 1980; Barahona *et al* 1982). In the former, the ground states of a quenched random Ising system have been studied by mapping into a suitable matching problem of graph theory. The latter paper also employs a similar strategy, and, being simpler to review, will be discussed here.

Barahona *et al* are interested in the ground state of a 2*D* Ising system (square lattice) with nearest-neighbour interactions. The exchange coupling J_{ij} has a probability x for assuming the value $-J$ and $(1-x)$ for taking on the value $+J$. Barahona *et al* noted that this problem could be mapped onto a problem in graph theory, known as the Chinese postman's problem. The latter has been solved, and the algorithm developed for this is now made use of to obtain results for the spin glass problem.

The Chinese postman's problem may be stated as follows: A postman delivers mail along a set of streets represented by the edges of a connected graph G. He must go along each street at least once in either direction. He starts from the post office (one of the nodes of G), and returns to it after delivering mail. If required, he may traverse the same street twice (see figure 30). The question is: What is the shortest possible route? The problem was posed by the Chinese mathematician Mei-Ko Kwan (1962) and solved by Edmonds (1965; see also Edmonds and Johnson 1973).

To proceed further, it is convenient to go back to the 2*D* spin glass problem posed earlier and cast it in a form suitable for our present purposes. Let us first introduce the definitions

$$W^{++}=\sum_{\langle ij\rangle} J_{ij} \text{ when } S_i=+1 \text{ and } S_j=+1$$

$$W^{--}=\sum_{\langle ij\rangle} J_{ij} \text{ when } S_i=-1 \text{ and } S_j=-1$$

$$W^{+-}=\sum_{\langle ij\rangle} J_{ij} \text{ when } S_i=\pm 1 \text{ and } S_j=\mp 1. \tag{58}$$

The Hamiltonian

$$\mathscr{H} = -\sum_{\langle ij \rangle} J_{ij} S_i S_j \tag{59}$$

can then be written

$$\mathscr{H} = K - 2W^{+-} \tag{60}$$

where

$$K = W^{++} + W^{--} + W^{+-} \tag{61}$$

and is a constant. From (60) we notice that finding the ground state of $\mathscr{H}$ reduces to minimizing W^{+-}. This problem is known in graph theory as the problem of finding a *minimum weighted cut.* The Chinese Postman's problem assumes a similar form, whence our interest in it.

Contact must now be made with graph theory, for which the following example is useful. Consider a triangle ABC with Ising spins at the vertices. The values of J_{ij} are taken to be as in figure 31. We now regard ABC as a graph G and construct its dual G′, first representing the regions of G by nodes as in figure 30(b). The dual graph G′ is now completed by drawing the appropriate bonds i.e., by connecting the nodes of G′ such that there is one bond (ij) in G′ for every edge (ij) of G.

Now there are $2^3 = 8$ spin configurations possible in the G we have considered, some of which are shown in figure 32, together with the values of $\mathscr{H}$. Declaring an edge of G to be a *cut* if it connects anti parallel spins, we also show in the figure the cuts corresponding to the various cases. A *quasi cycle* is that part of G′ relevant to the cuts. These are also shown. The central item of interest is the *violated edge* which is an edge ij of G for which the spin alignment is not consistent with that demanded by the sign of J_{ij}. The violated edges are also shown in figure 32.

Focussing now on the quasi cycle, for a cycle Q, a bond ij of Q is said to be violated if

$$J_{ij} > 0 \text{ and } ij \in Q$$
$$J_{ij} < 0 \text{ and } ij \notin Q.$$

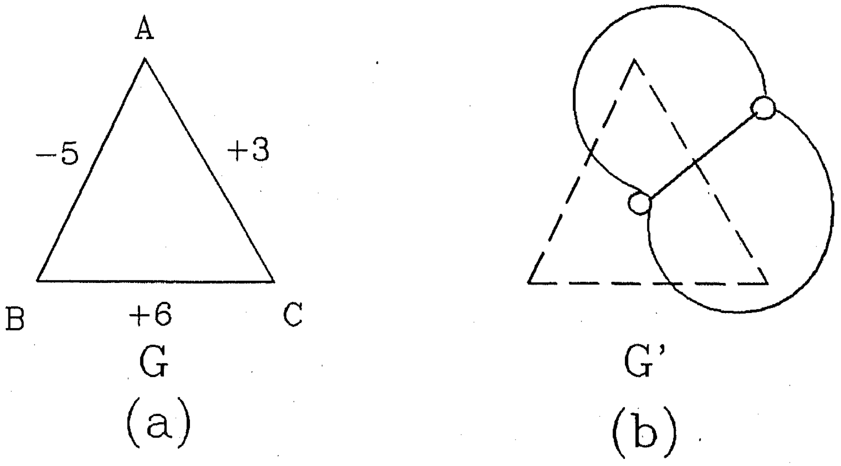


Figure 31. Three-spin problem in graph theory language. The Ising spins are at the vertices A, B, C of a graph G. The latter divides the plane into 2 regions, representative points of which are shown by the open circles. The dual graph G is constructed by linking the open circle as shown.

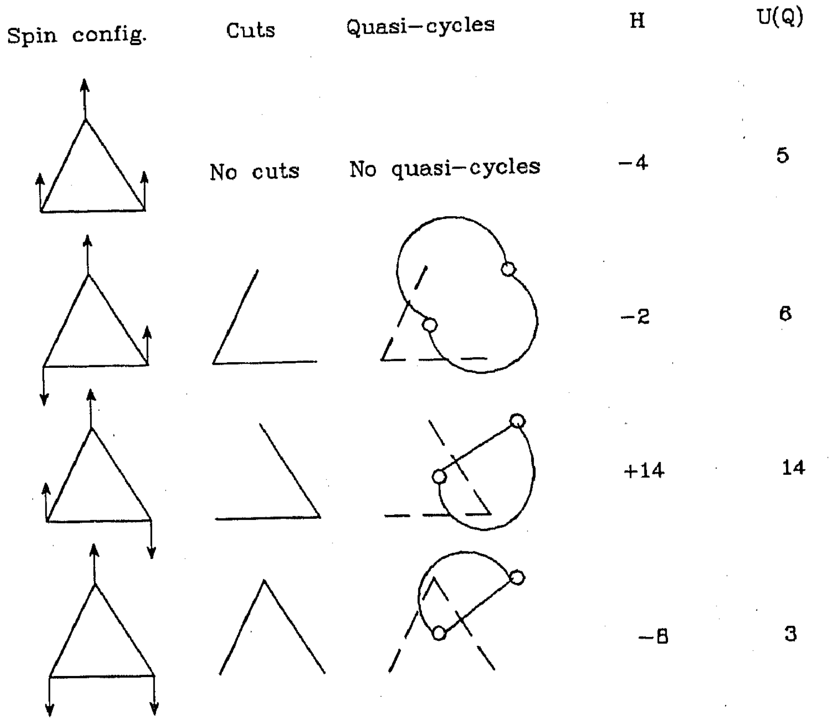


Figure 32. Graph-theoretic approach to the optimization of spin configuration. For explanations, see text.

Further, let

$$x_{ij}(Q) = 1 \text{ if } (ij) \text{ is violated}$$
$$= 0 \text{ otherwise.}$$

With these definitions, the problem of minimizing $\mathscr{H}$ can be shown to be equivalent to minimizing

$$U(Q) = \sum_{ij} |J_{ij}| x_{ij}. \tag{62}$$

In figure 32 the values of $U(Q)$ are given, and they are found to order the same way as do the values of $\mathscr{H}$ for the various spin configurations. In other words, minimizing $\mathscr{H}$ is the same as finding the minimum weighted quasi cycle in the dual graph G'. The latter problem can be solved by integer programming, as has been done by Edmonds. It is this method which has been exploited by Barahona *et al* for studying the ground state of the 2*D* Ising lattice. By analyzing the problem as a function of the parameter x (related to the probability of the exchange integral being positive), they have been able to highlight many interesting defect structures in the ground state.

5. Applications of simulated annealing

Besides TSP, simulated annealing has been applied to several other optimization problems, of which a sample will now be presented.

Earlier we discussed the drilling of holes in a PCB. Siarry and Dreyfus (1984) have

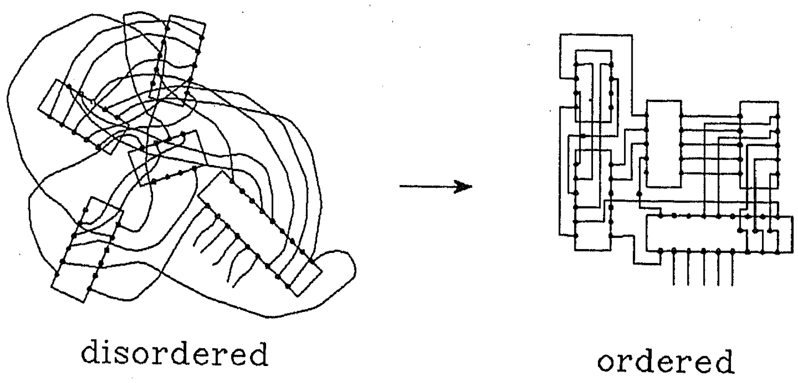


Figure 33. Layout of electronic components on a PCB, viewed as an order-disorder transformation.

attempted to use simulated annealing for the layout of the PCB. Traditionally, such layout is done using ECAD (electronic computer-aided design). However, these methods are wanting in some respects, which provided the motivation for the work by Siarry and Dreyfus. These authors view the problem as a disorder-order transformation (see figure 33). The components are treated as particles, and a suitable energy function is defined which is dependent on the lengths of the connecting strips and/or the crossing count. The high-temperature phase corresponds to a “liquid” of components while the low-temperature one corresponds to the properly laid out PCB wherein two electrically independent connections must not cross.

A one-shot phase transition as described above and leading to the desired layout remains at present an ideal; what has presently been achieved is a two-step design procedure which is conventional but different to the extent simulated annealing has been built into each step. In the first of these two steps, the components are dispersed across the PCB area in some sort of an optimal way, while the second step involves making actual connections between the various components via short, simple and straight paths. With proper routing, the total wire length would be a minimum.

Figure 34 illustrates how this works. There are twelve components, and these must be organized in a 3 × 4 array such that the total interconnection wirelength is a minimum. One starts with an arbitrary initial configuration. The Manhattan wire length L between two points (x_1, y_1) and (x_2, y_2) of the system is defined as

$$L = |x_2 - x_1| + |y_2 - y_1|. \tag{63}$$

The energy of the system is defined as an average Manhatten length of all the N wires i.e.

$$E = (1/N) \sum_{i=1}^{N} L_i. \tag{64}$$

The aim now is to make E as small as possible which, naturally, helps in keeping down communication delays. It is in optimising E that simulated annealing is used. The computational effort required is measured by the number of linkage configurations explored. Whereas the total number possible is around 5×10^8, the minimum in figure 34 could be reached in about $2{\cdot}2 \times 10^4$ attempts.

Some of the results obtained by Siarry and Dreyfus are shown in figure 35. Starting at a high temperature, the system is slowly cooled to the working temperature T.

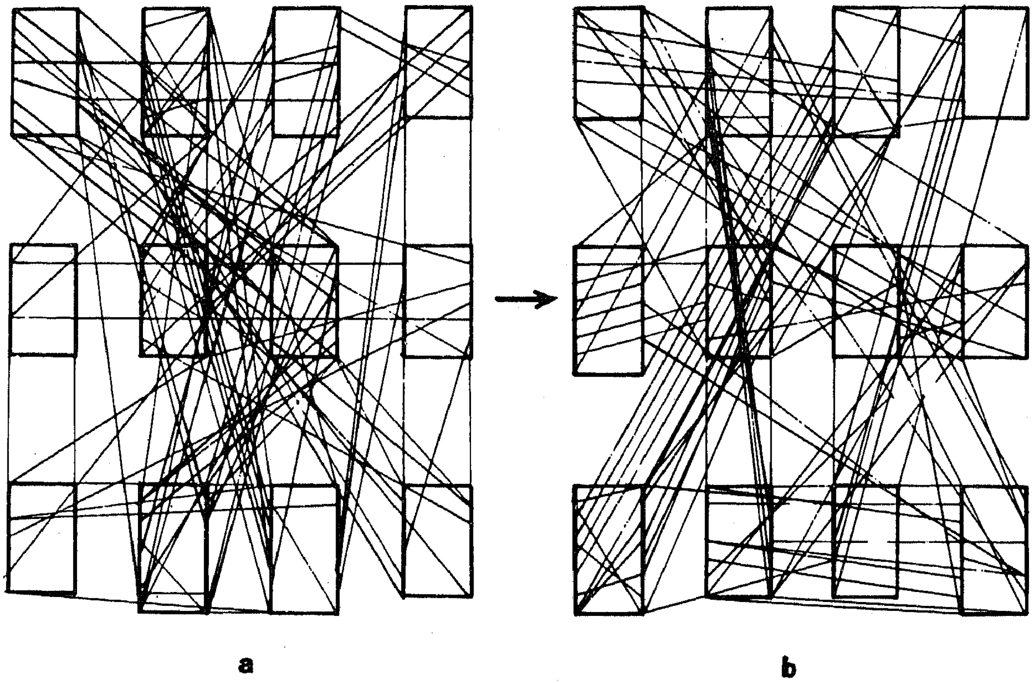


Figure 34. Placement of twelve components by minimizing the wire length. (a) shows the initial placement of the 12 components in an array. The energy corresponding to this is $E = 469$. After optimization, the configuration which emerges is shown in (b). Its energy is about half that of (a) (after Siarry and Dreyfus 1984).

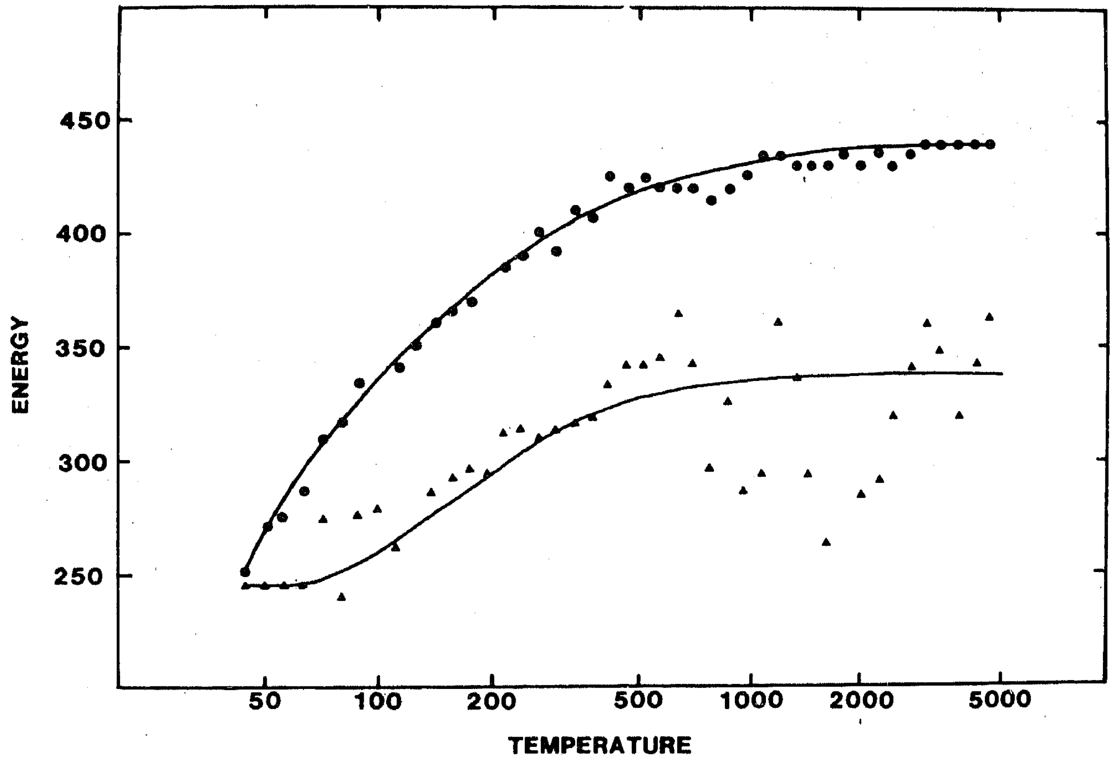


Figure 35. Average energy (dots) and minimum energy (arrows) assoicated with circuit layout, as a function of temperature (after Siarry and Dreyfus 1984).

The average of several trials as also the "minimum" observed at each temperature are plotted as a function of T. As is to be expected, the two curves approach each other as T is lowered.

Kirkpatrick *et al* (1983) at IBM have studied a whole gamut of problems related to circuit design, of which PCB layout is just one. While planning a complete system e.g. a computer, one first lays down the specifications, does the logic design and then the physical circuit design. Parts of the circuits would have to be compacted into chips and custom built, (apart from the standard, off-the-shelf chips used in the circuit design), chips then organized onto boards, boards subsequently arranged in layers and interconnected, and so on. Kirkpatrick *et al* show that at each stage there is some optimization involved, against a background of conflict of interests. After considering several illustrative examples of the various problems, they stress (for the first time) that in optimization problems with built-in frustration, there are always several, "near-optimal" solutions, all equally good and acceptable; therefore, it is not very fruitful to search for the notional, absolute minimum.

Several applications of simulated annealing have also been reported from the Los Alamos National Laboratory (Viswanathan *et al* 1985 and references cited therein), particularly in the context of the design of complex optical systems. As in the other cases, one supposes here that a given lens is in some kind of an energy state. The lower the energy, the better the lens i.e. the less aberrations it has. What one is looking for is the lens system corresponding to the lowest possible energy i.e. having minimum aberrations.

As in all engineering design, the design of an optical system start with target specifications. One then attempts to evolve a configuration of the system as a whole (which, naturally, would involve several components/subsystems) which would perform as closely as possible to the originally laid down specifications. Design is thus a mapping from specification space to system-configuration space. The mapping is defined by the optimization of an objective cost function, but unfortunately, for complex systems, one does not know *a priori* this cost function. In fact, a feel for it emerges only by going through several design exercises i.e. by going up a learning curve. Incidentally, design evaluation i.e. evaluating the performance of an already-designed system, is the converse of design mapping, and is better defined.

The Los Alamos optics group is working towards an intelligent lens design program. It will have two parts, one dealing with learning from previous experience and the other for performing an optimization using the best available cost function. It is in the latter that the simulated annealing technique is used. Actually a slight variant of the usual recipe is employed (Bohachevsky *et al* 1986) i.e. the Boltzman probability is modified from $P = \exp(-\Delta f/T)$ to $P = \exp(-\Delta f \cdot |f|^g/T)$ where T is the temperature and Δf the change in the original cost function f, consequent to a small displacement. The quantity g is an arbitrary, negative number. The above modification effectively decreases the probability of accepting a detrimental step, as the random walk approaches the global minimum (assumed to be zero). The new algorithm is referred to as *generalized simulated annealing*.

As application of interest to users of synchrotron radiation has been reported from Stanford University (Cox and Youngman 1985). This concerns the design of a device called the *undulator* in which the electron traverses a wiggly path, and in the process contributes to enhanced luminosity of the emitted radiation. The undulatory motion is effected by a periodic array of (SmCo) permanent magnets. The problem is that the magnets (numbering 240) are not identical in characteristics, and the question is how should they be organised so that the electron trajectory through the undulator is as close to the ideal as possible. Electron trajectory control is crucial, since even a few percent deviation from the ideal can cause a loss of the photon beam. The problem has been tackled by defining a suitable cost function and optimizing the design to minimize the cost, so that minimum cost corresponds to minimum distortion of the electron trajectory with respect to the ideal. The brute force approach of course is to try out *all* the possible magnet configurations which, however, is impossible since there are $\sim 10^{500}$ possibilities to be scanned. Simulated annealing was therefore tried, and it gave acceptable results after 1·5 million tries lasting ~ 3 hours on a VAX 11/780.

Mention may also be made of a war game application (Bohachevsky *et al* 1987) in which an intercontinental ballistic missile attack is assumed to take place and the defender has to deploy his protection resources so as to maximize the survival of his assets. The authors remark that evaluations could be performed in ~ 5 sec in a "high performance general-purpose computer. Such a time is sufficiently short to provide real-time assistance in battle management." We cite this example merely to point out that simulated annealing seems to be useful both in peace and in war!

6. Optimization strategies from the biological world

Biological evolution provides perhaps the ultimate example of a complex optimization problem. In the biological language, the Lin-algorithm may be described as a method in which one tries random mutations (alteration) of a trial solution and then selects the outcome if it is a fitter (cheaper) solution. This of course is done repeatedly. In the same language, simulated annealing may be described as a method which allows acceptance of unfavourable mutations with a finite probability. Independently, there has been interest in genetic strategies which address themselves to selecting the fittest species. Brady (1985) has applied such strategies to the TSP and reported improved performance over simulated annealing.

Brady found that the best performance was obtained with a method analogous to gene swapping during mating. Figure 36 illustrates what gene swapping means in our context. Here (a) and (b) show two tours, both containing the same segment (A . . . . . G), though the cities in the segment are visited in different order. Suppose the distance for the sequence ABCDEFG is smaller than for the sequence ACBDFEG. This means tour 2 can be shortened by replacing the segment ACBDFEG by ABCDEFG (see figure 37 (c)).

Brady has carried out several studies on a 64-city TSP. On a IBM 3081 D, he found that his strategy after 1 sec CPU time yielded a better answer than simulated annealing after 20,000 Monte Carlo steps (involving 5–9 sec of CPU time). He also notes that the bio-strategy may be ideally suited for parallel processing. Commenting on Brady's work, Bounds (1987) has remarked that there are not enough grounds as yet to conclude that simulated annealing is inferior to bio-strategies. Even if the latter appears to perform better for 64 cities, it is not certain that it would continue to do so for larger problems.

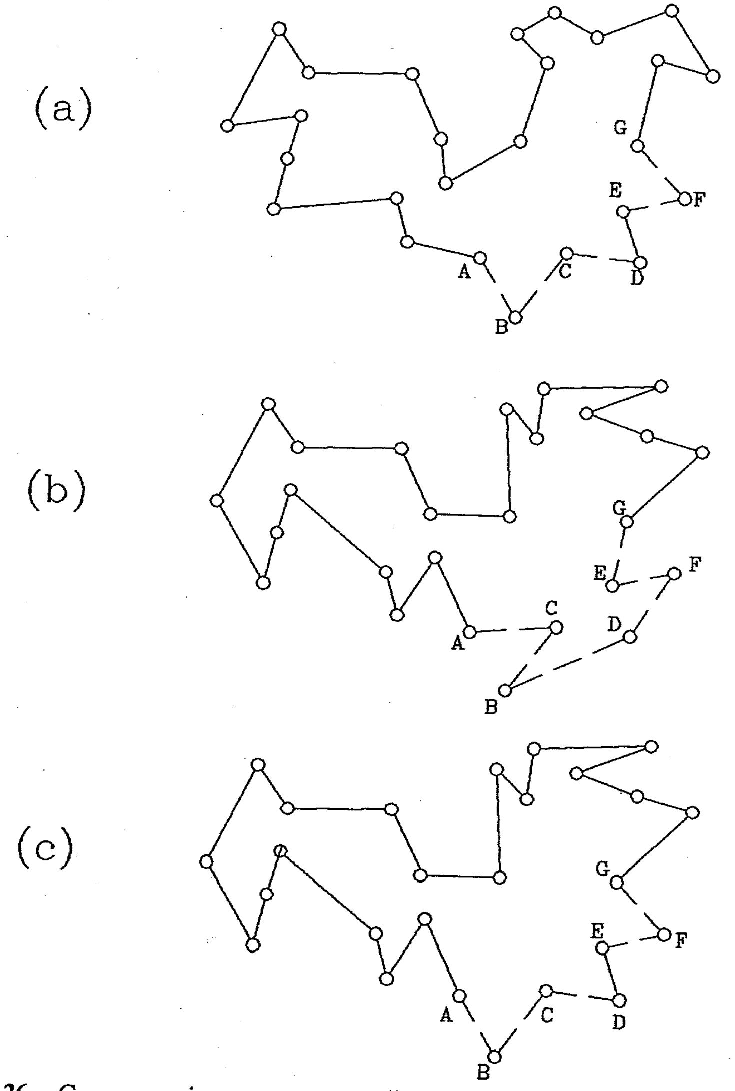


Figure 36. Gene-swapping concept as applied to the TSP. (a) and (b) show two tours (genes). Segment (A...G) occurs in both but the cities are visited in different order. The segment is shorter for (a) but the rest of the tour is better in (b). Tour (c) is produced by "gene swaping", and has the best of both (a) and (b).

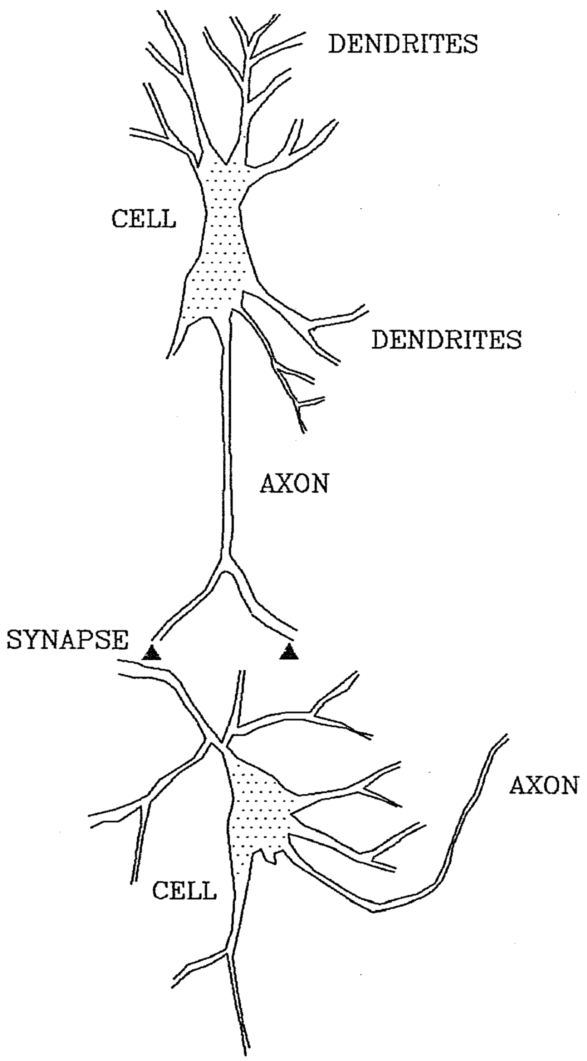


Figure 37. Sketch of the neuron and its environment.

7. Neural networks

7.1 *About neurons*

In the world of popular science, the term electronic brain is often used while referring to a computer. In reality, computers as we presently know them, work in an entirely different manner as compared to the human brain. While a modern supercomputer might execute with dazzling speed a numerical calculation that might take a human centuries to perform, it cannot, as yet, do some simple things which a three-year old can, like picking out a tree in a picture. True there are exciting developments like parallel computers, artificial intelligence etc., but even so, it will be a long time before the brain can be substantially mimicked.

Before proceeding further, some idea of the neuro-biological system is necessary. The brain consists of $\sim 10^{12}$ nerve cells or neurons. A highly schematic picture of a neuron is shown in figure 37 wherein three regions may be identified: (i) the cell (*soma*), (ii) the *dendrites* and (iii) the *axon*. The latter is the output channel linking to the other neurons. The neural network is quite complex, with each neuron having links with several thousand others through interfaces called *synapses*. In brief, a neuron receives inputs into its dendritic arbor from other neurons, and provides an input to other neurons via synapses on its axon. When a nerve cell "fires", an action potential (see figure 38) is produced. Whether a neuron fires or not is determined by the synaptic potentials which appear at the input end. Thus, neural activity is the result of the weighted integral of the activities of other cells, the weights being determined by the synaptic efficiencies. The neuron fires very rapidly emitting many pulses but it turns out that what is important is the frequency of firing or, equivalently, the *average* action potential.

A model is now needed which will not only represent all the above processes but

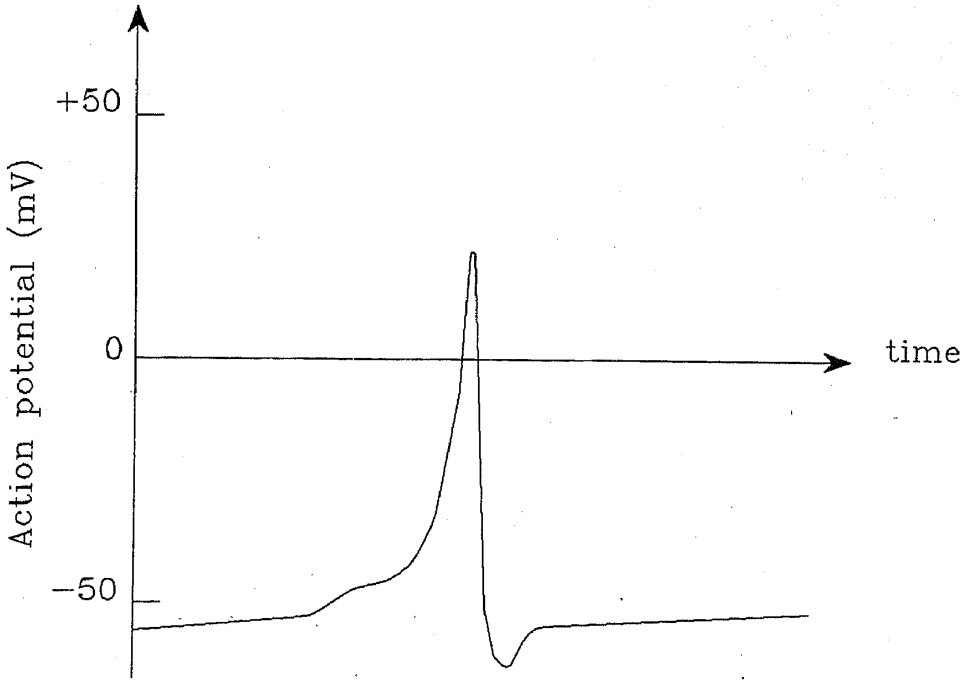


Figure 38. Profile of the action potential produced when a neuron fires.

will, in addition, represent some functions of the brain, like the recognition of elementary patterns. Three areas of intense activity relating to the modelling of neural networks pertain to (i) motor control, (ii) visual perception and (iii) pattern recognition. While the brain is able to accomplish all the above, modellists confine themselves (understandably) to one or the other aspect alone. Here only the last-mentioned of the above three will receive attention.

A practical system of interest in the context of pattern recognition is the content-addressible memory (CAM; see Tank and Hopfield 1987). This generally refers to a system wherein a pattern or a collection of them is already stored. When one of the patterns or even a corrupted version of it is presented to the system, it is recognized.

A CAM-type memory device must, among other things, have the following properties:

i) It must be able to learn.
ii) It must recognize a pattern previously learnt.
iii) It must be noise tolerant.
iv) It must be robust.

Figure 39 illustrates these attributes. Later we shall illustrate some of these features using models that have been proposed.

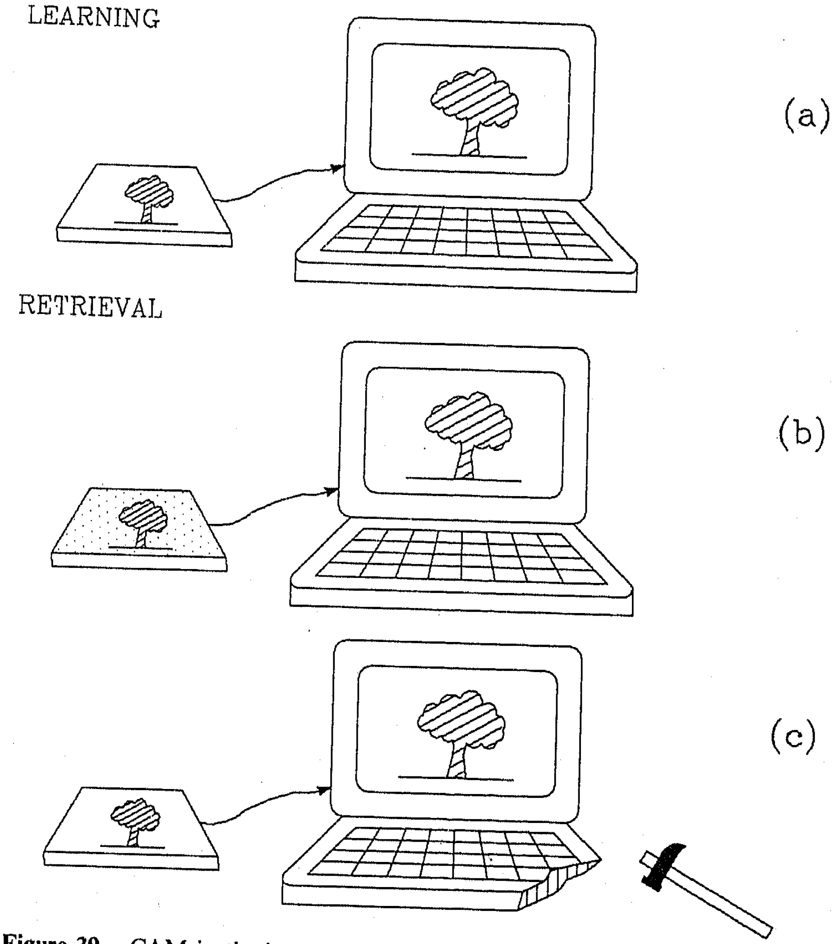


Figure 39. CAM in the learning mode (a) and in the retrieval mode (b), (c). (b) illustrates retrieval with a fuzzy input while (c) illustrates robustness.

7.2 *The Hopfield model*

From the early forties, electronic engineers have been interested in the neuron as a logical device. A neural network is an interconnection of such logical elements, with additional features to represent synaptic efficiency etc. The key element of course is the device itself, for which three different types of input-output characteristics have been proposed as in figure 40 (see, for example, Lippmann 1987). Of these, that in (a) is favoured by physicists while VLSI designers have preferred that in 40(c).

In 1982, Hopfield (1982) proposed a model of neural network with the basic element having an input-output characteristic (i.e. the average neuron potential) as in figure 40(a). The governing rules of the model are as follows:

i) The total input h_i to neuron i is given by

$$h_i = \sum_{j \neq i} T_{ij} V_j \tag{65}$$

where V_j is the output of neuron j. The element T_{ij} is representative of the synaptic connection between neurons i and j, and is assumed to be symmetric i.e. $T_{ij} = T_{ji}$.

ii) Neuron i responds to h_i according to the rules

$$\begin{aligned} V_i &= 1 \text{ if } h_i > 0 \\ V_i &= 0 \text{ if } h_i \leqslant 0. \end{aligned} \tag{66}$$

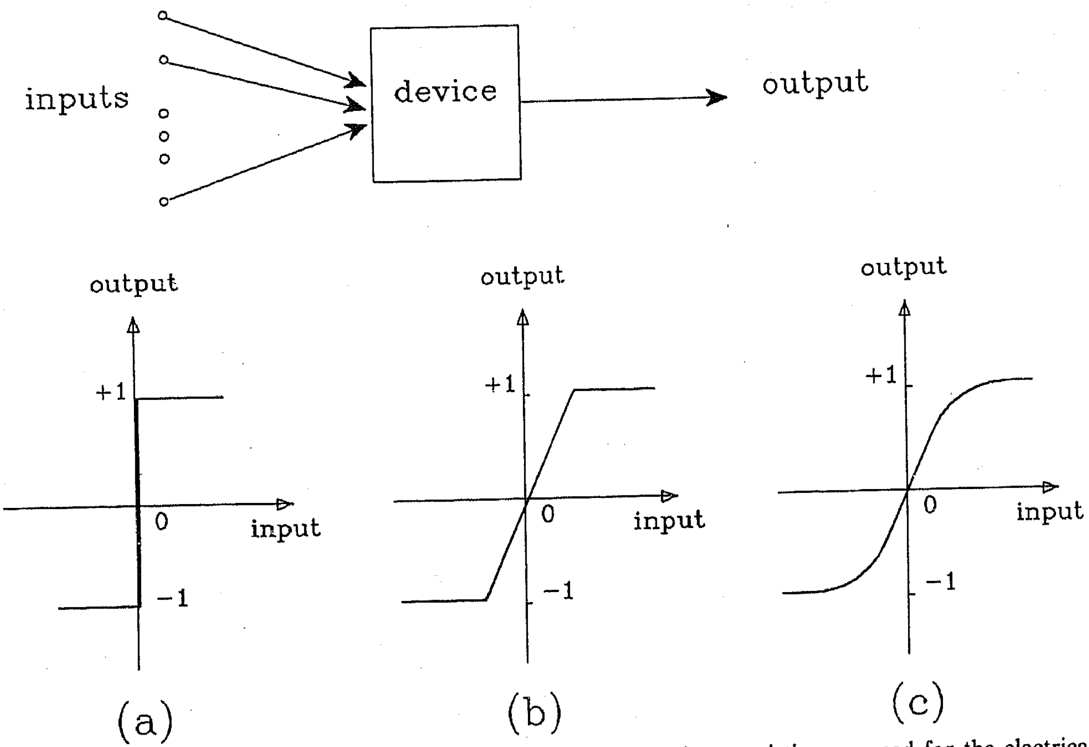


Figure 40. Three distinct types of input-output characteristics assumed for the electrical analogue of the neuron. The output denotes the average action potential.

The above rules are supplemented by the following evolution algorithm.

Step 0: Initialize as $V_i(t=0)=X_i$ say
Step 1: Calculate $\{V_i(t=1)\}$ as per rule (68).
Step 2: Continue, calculating $\{V_i(2)\}$, $\{V_i(3)\},\cdots$ etc., until $\{V_i(n+1)\}=\{V_i(n)\}$ i.e. a steady state is reached.

There is a one-to-one correspondence between the model described above and a lattice of Ising spins, as we shall see later. For the present we note that the steady states of the Hopfield model represent the minima associated with the Hamiltonian

$$\mathscr{H}=-\frac{1}{2}\sum_{i\neq j} T_{ij} V_i V_j. \tag{67}$$

In the Hopfield model, neural activity corresponds to the evolution of the neural circuit towards an appropriate minima, when suitably triggered by an input signal. Later we shall discuss some of the details, including storage and retrieval.

Hopfield soon moved away from this model to another where the neurons behaved as analog devices with characteristics as in figure 40(c) (Hopfield 1984). Let $g_i(u_i)$ denote the input-output characteristic (gain) of the neuron and $g_i^{-1}(V_i)$ the inverse. There are no longer step responses, and u_i lags behind the outputs V_j of the other cells because of input capacitances C_j and resistances R_j. The energy function is now given by

$$\mathscr{H}=-\frac{1}{2}\sum_{i\neq j} T_{ij} V_i V_j+\sum_i (1/R_i)\int_0^{V_i} g_i^{-1}(V)\,\mathrm{d}v \tag{68}$$

which is to be compared with (67). To understand the second term in (68), let us scale the gain g replacing

$$V_i=g_i(u_i) \text{ by } V_i=g_i(\lambda u_i)$$

and

$$u_i=g_i^{-1}(V_i) \text{ by } u_i=(1/\lambda)g_i^{-1}(V_i).$$

The variation of the gain curve with the scaling parameter λ is shown in figure 41. We observe that as $\lambda\to\infty$, the sigmoid curve approaches the step function and in this limit, the second term of (68) becomes negligible.

The difference between using (67) and (68) is highlighted in figure 42. Whereas for (67) the minima are located at the corners of the hypercube, in the case of (68) the minima could be in the interior. The ability to locate an attractor where one wants, could be a useful feature while designing memories.

Going back to (68), let us assume for simplicity $g_i=g$, $R_i=R$ and $C_i=C$, independent of i. The evolution equations then assume the form

$$(\mathrm{d}u_i/\mathrm{d}t)=-u_i/\tau+\sum_j T_{ij} V_j \tag{69}$$

with

$$\tau=RC \text{ and } V_j=g(u_j). \tag{70}$$

The above system works readily as a CAM. The desired patterns are learnt into the computer by choosing the set $\{T_{ij}\}$. The pattern presented to the system for

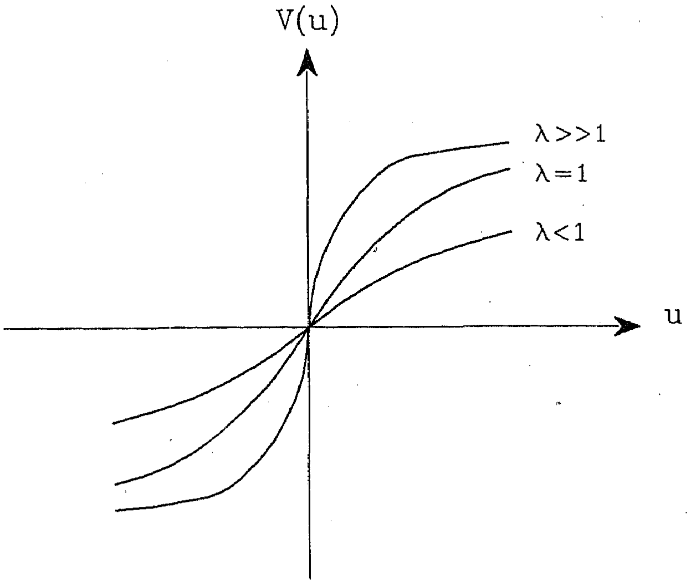


Figure 41. Variations in $V(k)$ the input-output characteristic, as the gain parameter λ is varied.

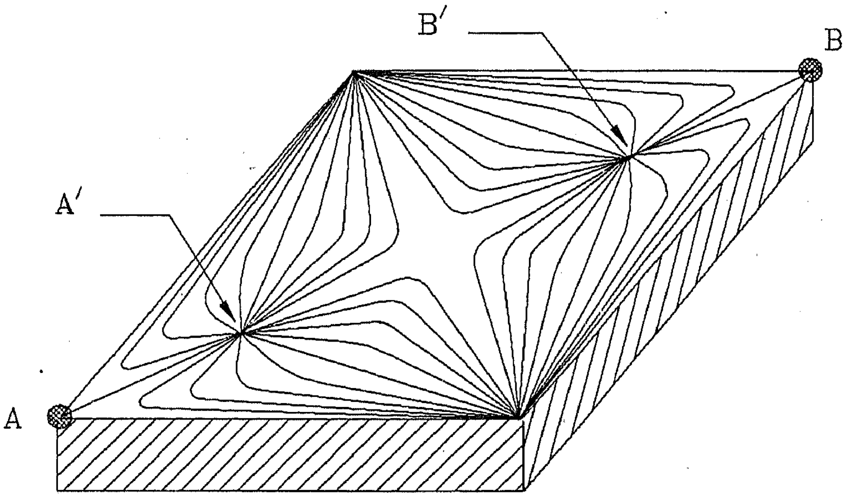


Figure 42. Attractors with discrete and continuous neurons. If the neurons behave as Ising spins, the attractors lie at the corners of an appropriate hypercube (in our illustration they are at A and B). For analog devices, the attractors can be inside, e.g. at A' and B'.

recognition is defined by the set of u_i's at time $t=0$. Given the T_{ij}'s and the set of u_i's at time $t=0$, the equations of motion provide a full description of the time evolution of the network. Basically, they drive the system towards one of the attractors (see figure 42).

7.3 *The TSP on a Hopfield network*

Hopfield and Tank (1985, 1986) have used a network as described above to study a 10-city TSP. To understand their work, let us consider a typical solution and represent it in matrix form as below.

$$
\begin{array}{ccccccccccccc}
 & & & & j\rightarrow & & & & & & & & \\
\mathrm{A} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & & \\
\mathrm{B} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & \\
\mathrm{C} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & & \\
\mathrm{D} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\
\mathrm{E} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & x & \\
\mathrm{F} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \downarrow & \\
\mathrm{G} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & & \\
\mathrm{H} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\
\mathrm{I} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\
\mathrm{J} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & &
\end{array}
$$

$$\text{TOUR: D H I F G E A J C B D} \qquad (71)$$

Following Hopefield and Tank, we write the matrix elements as V_{xj} where x labels the city and j denotes the column index. Matrix (71) has 100 elements, and if each of these can assume the values 1 or 0, then in all there are 2^{100} such matrices possible labelled by the 100-digit binary numbers ranging from $(0,0,\ldots 0)$ to $(1,1,\ldots 1)$. Just as the binary numbers (000), (001),... (111) can be associated to the corners of a three-dimensional cube (see figure 43), the 2^{100} binary numbers introduced above can be associated to the corners of a 100-dimensional hypercube. The matrix (71) (which has only one non-vanishing element along any row or along any column) and its various permutations form a subset of the hypercube corners.

Hopfield and Tank assigned the T_{ij}'s appropriate values for executing the TSP and represented the gain function by

$$V_{xi} = g(u_{xi}) = \tfrac{1}{2}(1 + \tanh(u_{xi}/u_0)) \quad \text{(for all } x, i) \qquad (72)$$

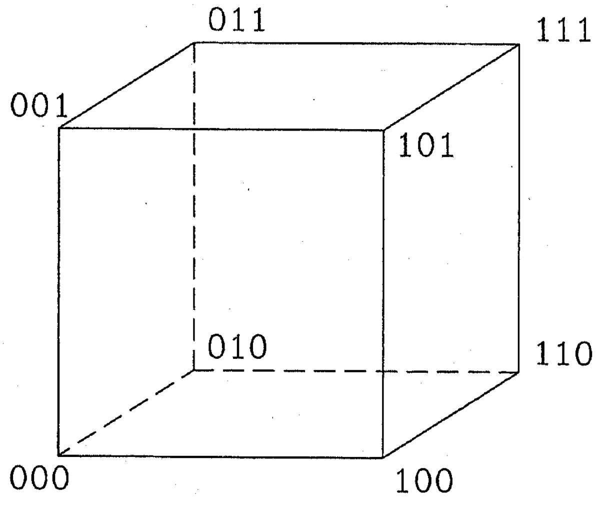


Figure 43. Labelling of cube corners using binary numbers.

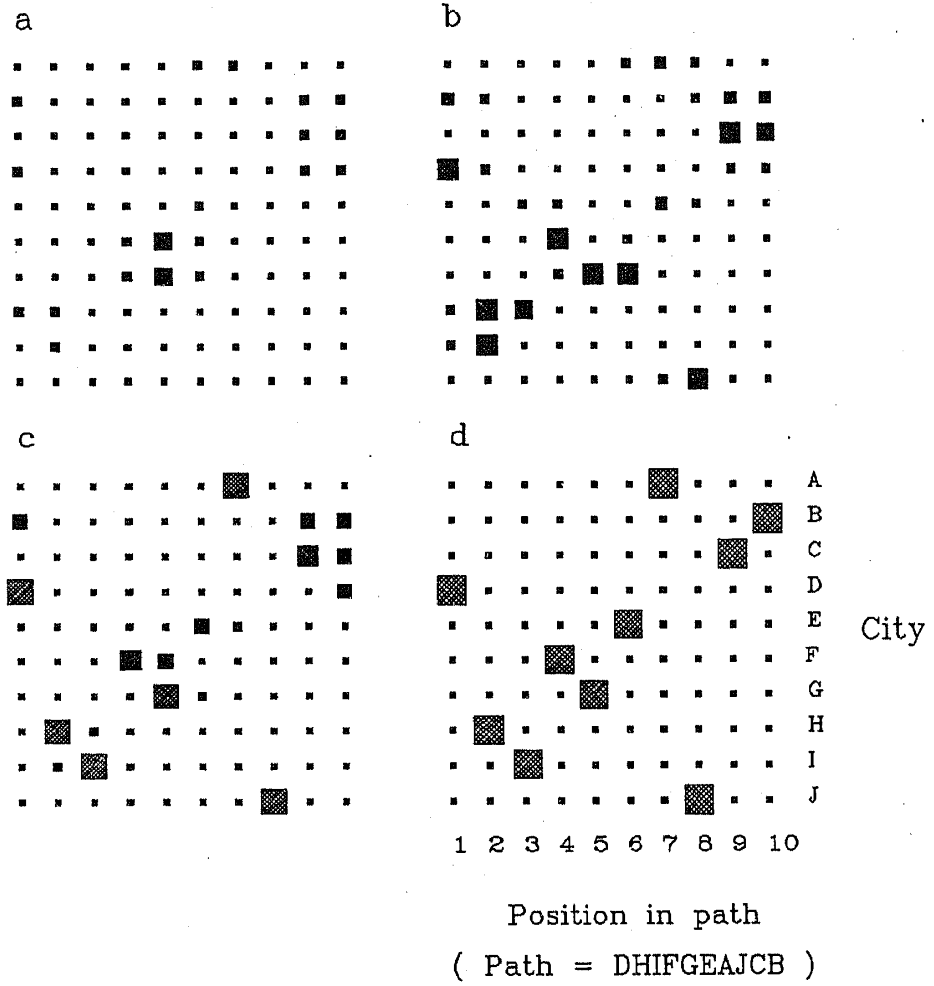


Figure 44. Analog Hopfield network at various stages during the solution of the TSP. For explanations, see text.

where u_0 is a constant (to be chosen suitably). The system is started off by setting

$$u_{xi} = \text{constant} + \delta u_{xi}$$

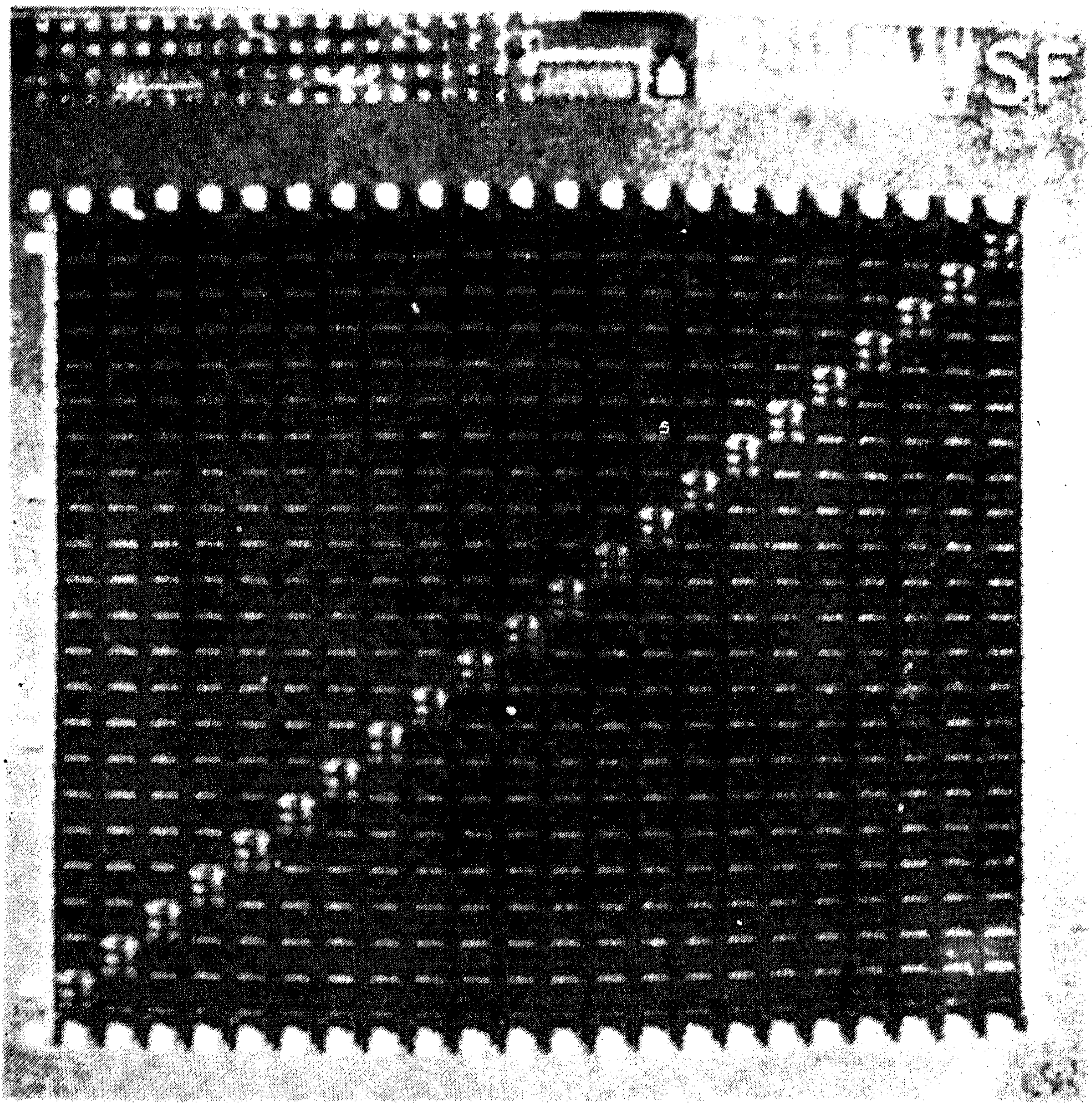
where δu_{xi} is randomly chosen in the interval

$$-0{\cdot}1\,u_0 \leqslant u_{xi} \leqslant 0{\cdot}1\,u_0.$$

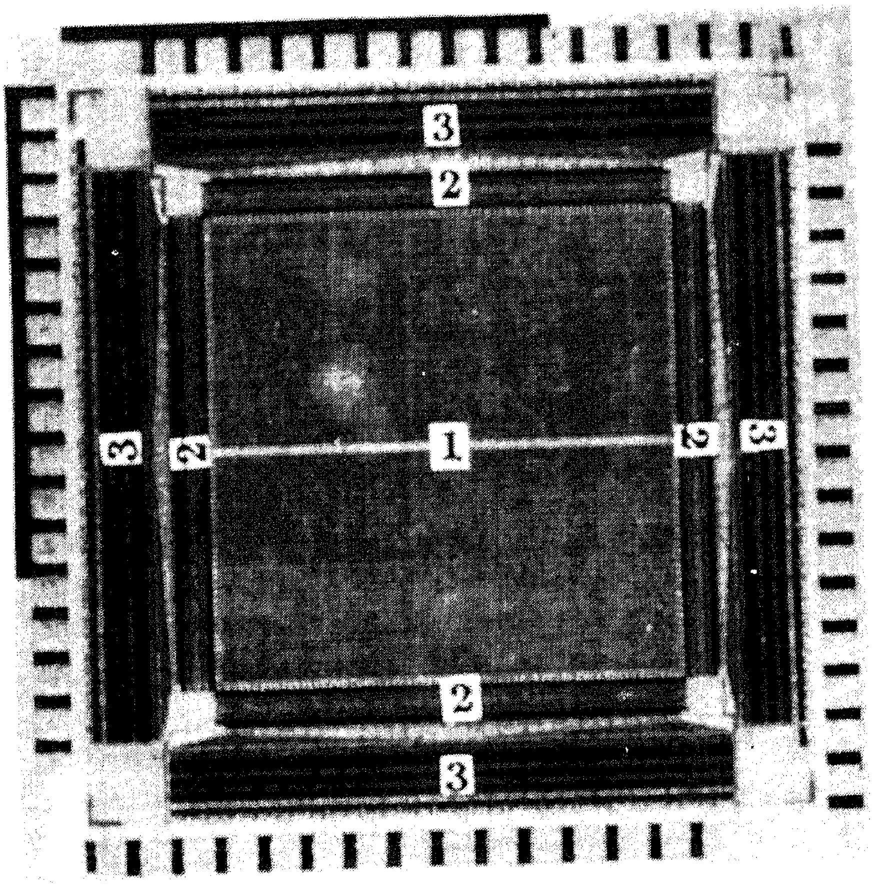
The system then evolves according to (69) until it reaches a steady state. Typical snapshots obtained during such a coast down are shown in figure 44, where the linear dimensions of the square are proportional to the output of the neurons in the array. The final result in (71) makes sense since only one neuron is "on" in each row, and the same is true of each column; such a situation represents a valid tour. With output patterns as in figure 44(a), (b) and (c) there is a problem. However, although a precise tour cannot be associated to such states, a qualitative interpretation is nevertheless possible. For example, in 44(c), row C has appreciable values for both the columns 9 and 10, although a large one for 9 than for 10. One interprets this situation by saying that at this stage, the network is still deciding whether city C should be in the 9th place or in the 10th. In other words, results for the intermediate stages are given a *probabilistic interpretation* by assigning relative probabilities in proportion to the sizes of the squares appearing in each row.

7.4 *VLSI implementation of neural network*

The early version of neural networks was built using discrete devices. A breakthrough occurred in 1986 when VLSI realizations of neural networks were reported for the first time (Sivilotti *et al* 1986; Graf *et al* 1986). At Caltech, Sivilotti *et al* designed a chip with 22 neurons (see figure 45). The size of the chip was around 6 mm × 6 mm and it was fabricated with 4-micron MOS technology. The authors have reported that a CAM with 289 neurons is under design. At AT & T Bell Labs, Graf *et al* have fabricated a 256 neuron chip, using a combination of analog and digital VLSI technologies plus a custom microfabrication process. Over a hundred thousand resistors, each with a value of a few megohms were necessary, and these could not be implemented with standard CMOS technology.



45(a)



45(b)

Figure 45. VLSI designed at Caltech (a) and at AT & T Bell Labs (b) (after Sivilotti *et al* 1986 and Graf *et al* 1986).

The Bell Labs chip is a candidate for use in an image processor. The circuit implemented has a fixed pattern of resistors and therefore the stable states are frozen in once the fabrication is done. A memory chip with programmable synapses has also been designed but it was found that the introduction of flexibility drastically reduced the number of neurons.

7.5 *Statistical mechanics and neural networks*

The study of neural networks from an electrical engineering point of view is several decades old. Interest of physicists in this subject is of more recent origin, dating back to the advent of the Hopfield model. The prime reason for this interest is the strong similarity (in some respects) to the spin glass, as also the ability to study some of the equilibrium properties of these networks via a statistical mechanical approach.

As already mentioned, the models of neural networks which physicists have been considering deal mostly with pattern recognition. The latter involves three stages namely, (i) learning, (ii) storage and (iii) retrieval given an input, which may even be partial or noisy. Of primary interest are (i) the storage capacity, and (ii) the ability to retrieve stored information with as little error as possible.

A fair amount of (theoretical) work has been done on these subjects during the last five years or so, and broadly speaking, they deal with either non hierarchical or hierarchical models. Each model centres around a particular learning scheme, and the objective is to achieve as large a storage capacity as possible, in addition to minimum corruption while retrieving. In some instances, the learning prescription has been chosen so as to mimic aspects of biological reality.

7.5.1 *The Hopfield model again*: We now go back to the Ising version of the Hopfield model i.e. (67) and, reverting to standard notation, write J_{ij} in place of T_{ij} and S_i, S_j instead of V_i, $\cdot V_j$. Understandably, this version has attracted much attention in the physics community.

In this model, the Ising spins play the role of neurons and the J_{ij}'s the role of the synapses. In the virgin system, all J_{ij}'s are equal to zero; this is the *tabula rasa* condition. Patterns are now learnt by the network using the so-called Hebb learning rule (Hebb 1949; Cooper 1973). A given pattern ζ^μ is described by the states of the N Ising spins, where N is the number of neurons in the network. Given the *tabula rasa* – type initial condition, the J_{ij}'s are obviously all equal to zero to start with. As patterns are learnt, the J_{ij}'s get progressively modified and in this way the image of the patterns is stored in the synapses. According to the Hebb rule, when the μth pattern is learnt, it modifies the (ij)th synapse by an amount $(\zeta_i^\mu \zeta_j^\mu)$. Thus, if p patterns are learnt, the synapse strength is given by

$$J_{ij} = \sum_{\mu=1}^{p} \zeta_i^\mu \zeta_j^\mu. \tag{73}$$

To retrieve a given pattern, one must feed an input corresponding either to complete or at least partial information concerning the pattern to be retrieved. One then expects the desired pattern to be retrieved by the dynamic evolution of the spin systems. However, the Ising model has no dynamics of its own i.e., the spins have no equations of motion. It is customary therefore to imagine that the spin system is in contact with a heat bath capable of flipping spins between the $+1$ and the -1 states (Glauber 1963). The spin state $\{S_i\}$ of the entire system now becomes time dependent, performing a random walk in the N-dimensional configuration space. The temperature of the heat bath is a measure of the stochastic noise in the dynamics. In the limit of zero temperature, the dynamics consists of every spin S_i flipping to align itself parallel to the local field $h_i(t)$ i.e.

$$S_i(t + \mathrm{d}t) = \operatorname{sign} h_i(t) = \operatorname{sign} \sum_{j \neq i} J_{ij} S_j(t). \tag{74}$$

Two possibilities now arise: (i) all the N spins are simultaneously updated according to (74), or (ii) one spin out of N (chosen randomly), is updated as per (74) in each step. The former scheme is referred to as parallel dynamics, and the latter as serial dynamics.

In the practical implementation of the Hopfield model, one first "teaches" the patterns to the Ising spin network which is done by setting J_{ij} according to the rule given in (73). The p patterns are now stored in the exchange couplings. To retrieve any one of this set, one presents an input set $\{S_i(0)\}$ to the network i.e., one adjusts the Ising spins to conform to the input pattern and then switches on Glauber dynamics. The network then evolves according to (74) until it reaches a fixed point wherein

$$S_i(t + \mathrm{d}t) = S_i(t), \quad \text{for all } i. \tag{75}$$

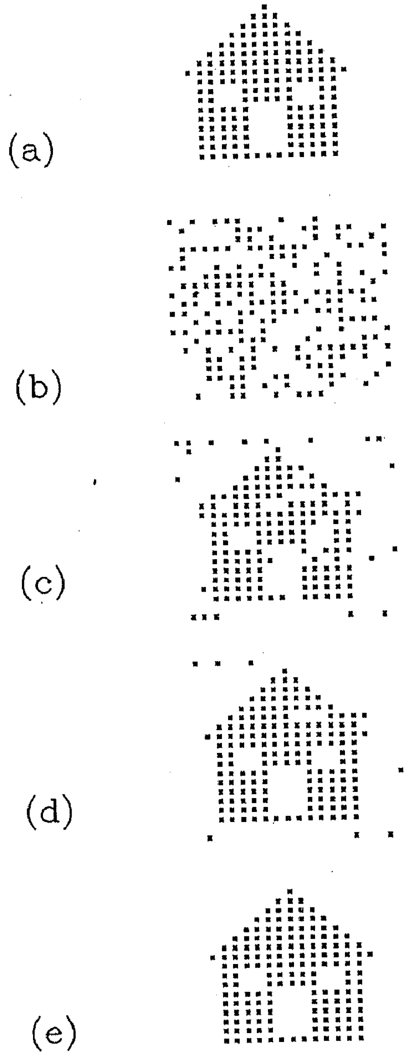


Figure 46. A (20 × 20) neural network of Ising spins was used for storing the pattern in (a). Nine other random-patterns were also stored. Subsequently, a highly corrupted pattern as in (b) was fed. Using a parallel dynamics algorithm, the system was able to retrieve the correct pattern in 4 steps.

The persisting pattern is the pattern retrieved. As Amit *et al* (1987) remark, "The basic goal of the synaptic modifications is to create attractors for subsequent dynamic processes, which are to retrieve learnt information".

Figure 46 provides an example of retrieval carried out in the above fashion. A pattern corresponding to a hut was stored in 20 × 20 network of Ising spins. Nine other random patterns were also stored, making $p = 10$. The pattern in figure 46(b) was fed as an input to the network, and the latter allowed to evolve according to (74). It is observed that the spin state of the system converges to the learnt pattern in 4 steps.

Amit *et al* (1985, 1987; see also Sompolinsky 1988) have explored in detail the statistical mechanics of the Hopfield model, particularly with a view to assess its storage and retrieval capabilities. Their results may be summarized as below. Firstly, Amit *et al* have noted that the Hopfield model is exactly soluble. Its equilibrium properties are succintly described by the phase diagram in figure 47. Here the x-axis

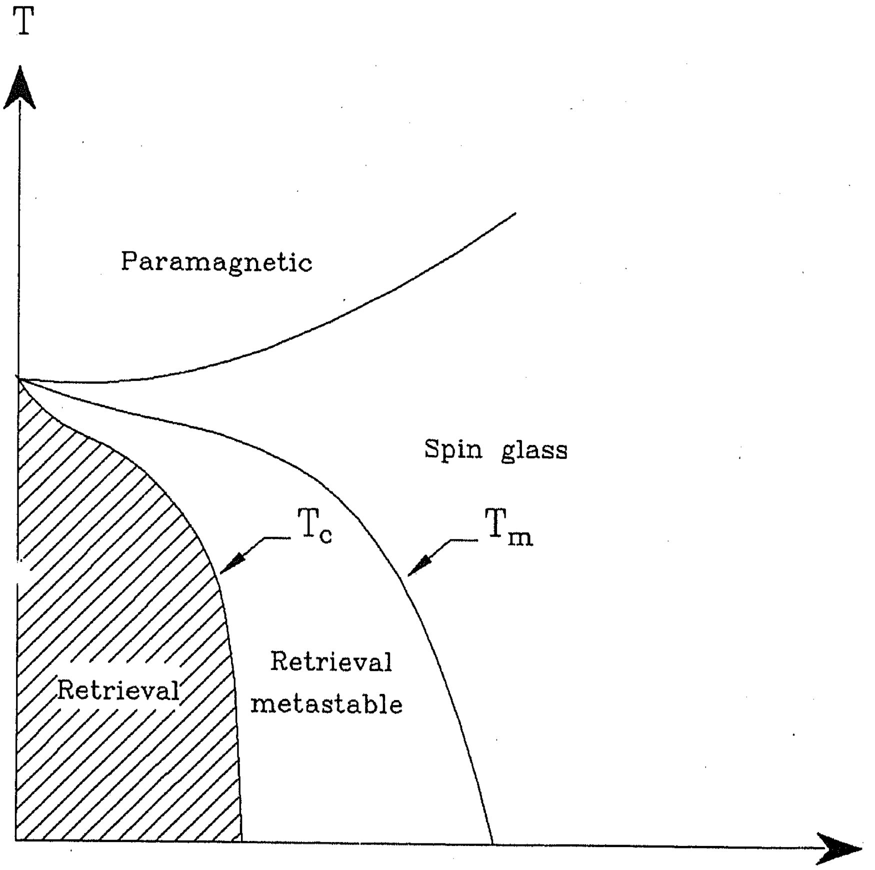


Figure 47. Phase diagram for the Hopfield model. For effective operation of the memory, the system must be within the shaded region. Further explanations are given in the text.

corresponds to a parameter $\alpha = p/N$. There are several phase boundaries, implying the existence of different phases. Of these, that marked *Retrieval* is an 'ordered' phase and of interest for operating the network as a CAM .

To understand the phase diagram, let us first consider the case $\alpha = 0$. This is the unsaturated case i.e. p is finite while $N \to \infty$. One wants to know whether there is a phase transition as T is decreased from a high value. To answer this question, Amit *et al* study the free energy density given by

$$-\beta f(\beta) = \lim_{N \to \infty} \{N^{-1}[\ln \mathrm{Tr} \exp(-\beta \mathscr{H})]_{\mathrm{Av}}\} \tag{76}$$

where $[\cdots]_{\mathrm{Av}}$ stands for an average over the stored patterns $\{\zeta_i^\mu\}$. It is a quenched average. The correlation of the thermodynamic states of the system with the embedded patterns are the order parameters. The order parameter $\mathbf{m}$ is a p-component vector with components

$$m^\mu = (1/N) \sum_i \langle S_i \rangle \zeta_i^\mu \tag{77}$$

where $\langle S_i \rangle$ is the *thermal* average of the spin at site i. Above $T = 1$, the system is in the paramagnetic phase with $\mathbf{m} = 0$ but below $T = 1$, states with non-zero $\mathbf{m}$ are possible. There are several types of these, of which the most important is of the form

$$\mathbf{m} = m(1, 0, 0, \ldots\ldots 0) \tag{78}$$

i.e. only the first entry is non-zero and all the $(p-1)$ remaining ones are zero. Obviously one could also have equivalent solutions obtained by permuting the position of 1 as also by replacing 1 with -1; in other words, there are in all $2p$ solutions of the type (78). These are called *Mattis states*, being related to the states known in a model of magnetism called the Mattis model. Besides the Mattis states, other symmetric solutions are possible having the form

$$\mathbf{m} = m(1, 1, \ldots\ldots 1, 0, 0, \ldots\ldots 0), \tag{79}$$

where the first n components are unity and the remaining $(p-n)$ are zero. Naturally, one must also count the equivalent solutions. In addition, asymmetric solutions like

$$\mathbf{m} = (3/8, 3/8, 1/4, 1/2, 1/4, 0, 0, \ldots\ldots 0) \tag{80}$$

are also possible.

Amit *et al* show that between $T=1$ and $T=0{\cdot}461$, Mattis states are the only ones possible. Below $T=0{\cdot}461$, symmetric states like (79) also appear and are dynamically stable. From the memory storage point of view they are not important since they represent a mixture of Mattis states (i.e. a mixture of stored patterns); in fact they are spurious states. At very low temperatures, some asymmetric states also appear which are metastable. Summarizing, one could say that for $\alpha=0$, storage is possible below $T=1$.

For achieving high storage capacity, one would naturally like to increase α. However, the case $\alpha \neq 0$ is slightly more complicated; but it is also interesting, as it has linkages with what was discussed earlier. Specifically, the replica method becomes necessary to handle the problem of configuration averaging, and one has to have several order parameters. Firstly there is the magnetization defined by

$$m_\rho^\nu = (1/N)\left[\sum_i \zeta_i^\nu \langle S_i^\rho \rangle\right]_{\mathrm{Av}}, \tag{81}$$

then the Edwards-Anderson parameter

$$q_{\rho\sigma} = (1/N)\left[\sum_i \langle S_i^\rho S_i^\sigma \rangle\right]_{\mathrm{Av}} \tag{82}$$

and finally a parameter $r_{\rho\sigma}$ related to the magnetic correlations $[m_\rho^\mu m_\sigma^\mu]_{\mathrm{Av}}$.

At this stage, one would have expected a Parisi kind of approach but Amit *et al* prefer to respect replica symmetry since breaking that symmetry brings about only marginal changes. Thus they suppose

$$\begin{aligned} m_\rho^\nu &= m^\nu \\ q_{\rho\sigma} &= q \quad \rho \neq \sigma \\ r_{\rho\sigma} &= r \quad \rho \neq \sigma, \end{aligned} \tag{83}$$

and study the mean field equations. The phase diagram shown in figure 47 is the outcome. Mattis-like states are once again possible, but with $\mathbf{m} < 1$. Such states are referred to as *retrieval* states. There are many phase boundaries in figure 47, and they may be understood as follows. In the region marked retrieval, the states are global minima. Between the lines T_c and T_m is a region where the retrieval states are metastable. The spin glass phase is where retrieval states are unstable. Lying above all these is the paramagnetic phase. T_c is a first-order transition, and so is T_m.

7.5.2 *Beyond the Hopfield model*: The Hopfield model has many limitations, a serious one being that correlated patterns are difficult to learn. To appreciate this, we first note that perfect retrieval implies that given an input pattern ζ^ν to a network wherein p patterns including ζ^ν are stored, the network retrieves ζ^ν via dynamical evolution. In turn, this implies that

$$\sum_{j\neq i} J_{ij}\zeta_j^\nu \text{ must equal } \zeta_j^\nu.$$

Let us check if this is true. Using (73) we have

$$\begin{aligned}\sum_{j\neq i} J_{ij}\zeta_j^\mu &= \frac{1}{N}\sum_{j\neq i}\sum_{\mu=1}^{p}\zeta_i^\mu\zeta_j^\mu\zeta_j^\nu\\ &\approx \frac{1}{N}\sum_{j=1}^{N}\zeta_i^\nu\zeta_j^\nu\zeta_j^\nu+\frac{1}{N}\sum_{\mu\neq\nu}\sum_{j=1}^{N}\zeta_i^\mu\zeta_j^\mu\zeta_j^\nu\\ &\approx \zeta_i^\nu+\sum_{\mu\neq\nu}\zeta_i^\mu\sum_{j=1}^{N}\frac{1}{N}(\zeta_j^\mu\zeta_j^\nu)\\ &\equiv \zeta_i^\nu+\sum_{\mu\neq\nu}\zeta_i^\mu C_{\mu\nu}(\text{say}).\end{aligned}\tag{84}$$

Here $C_{\mu\nu}$ is the matrix which describes the correlation or overlap between the patterns learnt by the network, and unless this overlap is identically equal to zero, the retrieved image would always be corrupted by some noise. While $C_{\mu\nu}$ vanishes in case the patterns ζ^μ and ζ^ν are orthogonal, in the case of random patterns it is $O(1/\sqrt{N})$. In practical applications, the stored patterns are seldom orthogonal or random.

Personnaz *et al* (1985; see also Kanter and Sompolinsky 1987) have suggested a way of overcoming the above difficulty. They observe that the Hebb learning rule is local in character i.e. each J_{ij} depends only on the values of ζ_i^μ and ζ_j^μ. Locality of learning certainly makes biological sense but, implemented in an artificial neural network, it poses unacceptable limitations. Personnaz *et al* propose getting over this difficulty by adopting a non local learning prescription. In this model,

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\sum_{\nu=1}^{p}\zeta_i^\mu\zeta_j^\nu(C^{-1})_{\mu\nu}\tag{85}$$

where C^{-1} is the inverse of the correlation matrix introduced in (84). To see how the above learning prescription facilitates cleaner retrieval than via (73), we switch to matrix notation and first write

$$C=\frac{1}{N}\zeta^T\zeta \text{ and } J=\frac{1}{N}\zeta C^{-1}\zeta^T\tag{86}$$

where ζ is the matrix whose elements are the numbers ζ_i^μ and ζ^T is the transpose of

ζ. With the above notation we have

$$J\zeta = \frac{1}{N}\zeta C^{-1}\zeta^T\zeta \doteq \zeta C^{-1}C = \zeta, \tag{87}$$

showing perfect retrieval.

Learning rule (85) implies that the matrix C^{-1} exists, which in turn implies that the patterns ζ^{μ}'s are linearly independent.

One of the desirable qualities of a neural network is perfect memory. This implies that given an input pattern that has some resemblance to one of the memorized patterns, the output is the correct and uncorrupted version of the concerned pattern. It turns out that in the case of the Hopfield model together with the Hebb learning rule, if the input is not sufficiently near to one of the memorized patterns, the network can get so confused as to get into an asymptotic state which is far removed from every memorized pattern. Parisi (1986) has a learning prescription which prevents the output being in a state of total confusion. Basically, like in the Hebb rule, J_{ij} is modified everytime a new pattern is learnt. However, unlike in (73), one has

$$J_{ij}^{\text{new}} = g(J_{ij}^{\text{old}} + \zeta_i^{\mu+1}\zeta_j^{\mu+1}) \tag{88}$$

where g is an appropriate nonlinear function such that J cannot become arbitrarily large (or small) i.e. remains within bounds. What happens with (88) is that old memories fade away i.e. only the last αN patterns are stored and the remaining ones are forgotten. Thus storage beyond capacity (αN) is avoided, and so also is confusion.

Gardner (1988) also addresses the question of increasing the storage capacity. Her point is that successful retrieval depends not merely on the existence of fixed points, but also on the basins of attraction. A parameter k introduced by her dictates the basin size; the larger the value of k, the greater is the basin size.

In her scheme, learning is not by a formula but by a small algorithm. Let us assume that one starts with a network containing no patterns at all, and wishes to store a pattern ζ^{μ}. Gardner's algorithm has the following steps.

(1) Start with arbitrary values for J_{ij}'s, subject however to the constraints

$$\text{(a) } J_{ii} = 0, \text{ for all } i; \quad \text{(b) } \sum_{j\neq i} J_{ij}^2 = N, \text{ for all } i. \tag{89}$$

(2) Choose a positive value for k and define a mask

$$\varepsilon_i^{\mu} = \theta\left[k\left(\sum_{j\neq i} J_{ij}^2\right)^{1/2} - \sum_{j\neq i} J_{ij}\zeta_i^{\mu}\zeta_j^{\mu}\right], \tag{90}$$

where $\theta[x]$ is the Heaviside step function.

(3) Compute

$$\Delta J_{ij}^{\mu} = \varepsilon_i^{\mu}\zeta_i^{\mu}\zeta_j^{\mu}. \tag{91}$$

(4) Next compute

$$J_{ij}^{\text{new}} = J_{ij}^{\text{old}} + \Delta J_{ij}^{\mu}. \tag{92}$$

(5) Using the value of J_{ij} as given by (92), go back to step (2) and carry through steps (2)–(5) till ε_i^μ vanishes for all i.

Pattern ζ^μ is now stored. Other patterns can be added in a similar fashion. It may be noted that when ε_i^μ vanishes, we have

$$\sum_{j \neq i} J_{ij}\zeta_i^\mu \zeta_j^\mu > k. \tag{93}$$

Gardner proves that provided (93) is satisfied, there exists a fixed point for the pattern ζ^μ.

We now turn to the physical significance of k. By using the replica method, Gardner studies the dependence of α_c the maximum storage capacity on k. The qualitative behaviour is sketched in figure 48, and it highlights how α_c depends on the correlation between the patterns. For simplicity, it is assumed that all patterns have the same magnetization **m**. Perfect correlation corresponds to $\mathbf{m} = 1$ and no correlation at all to the case $\mathbf{m} = 0$. For sure retrieval, obviously the basin i.e. k must be large. As can be seen from the figure, k can be made large only by paying a price i.e. by reducing α_c the storage capacity. For small k, Gardner's prescription seems to do better, the greater is the correlation between patterns.

As already remarked, the initial state in the Hopfield model is a *tabular rasa.* If one takes this seriously vis-a-vis biological neurons, then the organization of the synapses in an adult is the direct result of learning or as Toulouse *et al* (1986) refer to it, the result of "direct instructive prints by the environment." This seems a bit far fetched, and it has been argued that the initial state of the synapses is already highly structured and that subsequent evolution in the adult is mainly the result of suitable modifications. In other words, learning is by pruning and selection.

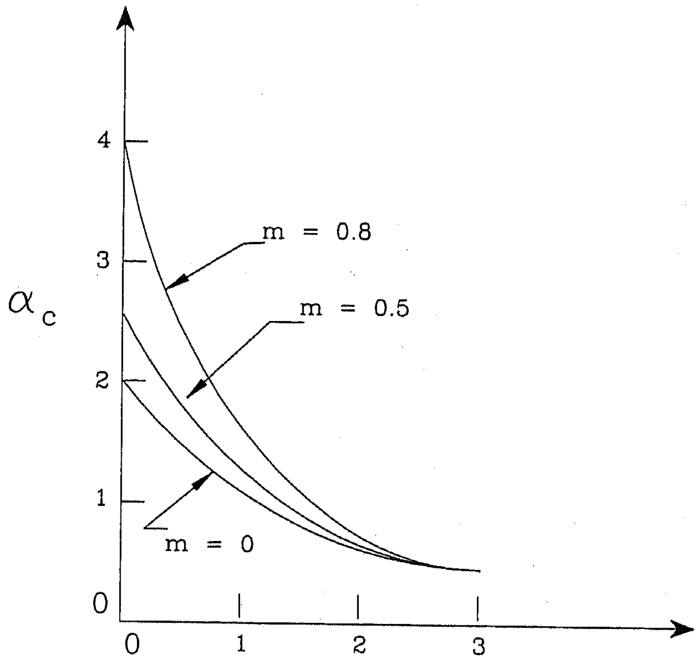


Figure 48. Schematic plot of the relationship between storage capacity α_c and the basin size k, as deduced by Gardner.

Toulouse *et al* seek to describe learning by selection, by analogy to the spin glass. An energy function as in (67) is introduced but, unlike in the Hopfield model where the cost landscape associated with the initial state is flat and featureless, the landscape is now taken to be a complex one with valleys within valleys etc., as in the case of the spin glass. Obviously this implies that the starting J_{ij} are random, and indeed Toulouse *et al* assume that the synapses have initial strengths of either $+J$ or $-J$ with a random distribution.

Suppose a pattern is fed to such a system. The initial spin configuration would then evolve via dynamics towards an attractor i.e. towards the bottom of one of the valleys already existing. In other words, the spin glass-like terrain offers scope for storage. However, it might not be particularly useful in an "as is" condition but it could conceivably made so by suitably modifying the synaptic strengths. Thus, Toulouse *et al* suggest that the storage of pattern ζ^μ could by represented by the change

$$\Delta J_{ij}^{(\mu)} = \frac{\varepsilon J}{\sqrt{N}} \zeta_i^\mu \zeta_j^\mu. \tag{94}$$

The effect of the change (94) is to alter slightly the landscape, at least in so far as the valley where the pattern ζ^μ rests. The basin of attraction is enlarged, probably at the expense of other neighbouring valleys. Every pattern learnt causes a change *a la* (94) and the synaptic strength thus undergoes a random walk with steps of length $\sim (\varepsilon J/\sqrt{N})$. Concomitantly, the whole landscape evolves during the learning process. Thus, starting from a hierarchical distribution of valleys, learning can be likened to the pruning of a tree. Observe that in this scheme, the already stored information influences the next event (in terms of the room available for the basin). Thus, a pattern that is repeatedly learnt, has better retention possibilities compared to one learnt say just once.

Toulouse *et al* mention that they have carried out numerical studies to establish the salient features of their model. No detailed results are reported. However, they mention that for $\varepsilon \geqslant 2{\cdot}5$ retrieval quality was perfect.

We turn now to hierarchical models (Dotsenko 1985; Parga and Virasoro 1986; Gidas 1989). Among these, we can distinguish two broad categories: (i) those in which information is stored and processed (during retrieval) in a hierarchical fashion, and those in which the patterns stored are themselves hierarchial in character. We consider first an example belonging to the first category (Gidas 1989).

Gidas does not consider neural networks as such but a problem in image processing. As he describes it, his method "is based on a combination of renormalization group ideas, the Markov random field modelling of images, and Metropolis-type Monte Carlo algorithms. The method is efficiently implementable on parallel architectures, and provides a unifying procedure for performing hierarchial, multiscale, coarse-to-fine analysis of image processing tasks such as restoration, texture analysis, coding, motion analysis etc." Here we shall present to highlight the above model in our notation.

Basically, learning i.e. storage is done at a hierarchy of levels; naturally, retrieval is also likewise. Consider a pattern ζ^μ that is to be learnt. Let there be $(l+1)$ levels $0, 1, 2, \ldots . l$ associated with the learning process so that $\zeta^\mu (\equiv \zeta^{(0)\mu})$ goes through the sequence

$$\zeta^{(0)\mu} \xrightarrow{p_0} \zeta^{(1)\mu} \xrightarrow{p_1} \zeta^{(2)\mu} \xrightarrow{p_2} \cdots \xrightarrow{\quad p_{(l-1)} \quad} \zeta^{(1)\mu}. \tag{95}$$

Each step p_i in the above involves a coarsening procedure or what Gidas refers to as a *renormalization transformation.* In each step, there is a reduction in resolution by a factor k, say. If there are N lattice sites in $\zeta^{(0)\mu}$, there would be (n/k) in $\zeta^{(1)\mu}$ and so on. Thus $l=\log_k N$ where k is the scale of reduction. The sequence (95) converts a given pattern into $(l+1)$ patterns, and that at the mth level is learnt according to the rule

$$J_{ij}^{(m)\mu}=\{1/k^{(l-m)}\}\cdot\zeta_i^{(m)\mu}\zeta_j^{(m)\mu} \tag{96}$$

In this way, patterns can be learnt and stored at a sequence of levels. As for the coarsening procedure, many possibilities exist and Gidas discusses a few.

Turning now to retrieval, consider an input $\sigma^{(0)}$, which could even be noisy. This too is processed through various levels, the lth one being given by

$$\sigma^{(l)}=p_{l-1}p_{l-2}\cdots p_0\sigma^{(0)} \tag{97}$$

where $p_0, p_1, \ldots$, etc. refer to the successive coarsening transformations. The lth level input $\sigma^{(l)}$ is used to retrieve from among the images stored at the same level. Let the retrieved pattern be denoted $\Sigma^{(l)}$. By a suitable inverse transformation p^*, one now generates $\Sigma^{(l-1)}$ i.e.

$$\Sigma^{(l-1)}=p_{l-1}^*\Sigma^{(l)}.$$

In this way, by successively applying the inverse transformation, one recovers the image $\Sigma^{(0)}$. Under ideal circumstances, this would correspond to the uncontaminated pattern whose noisy representative is $\sigma^{(0)}$. It should be mentioned that issues like storage capacity, ability to store correlated images, corruption in the retrieved images etc. are yet to be studied.

Parga and Virasoro (1986) also consider encoding to be implemented in structured layers. Their idea is that such a learning/storage scheme is advantageous when the patterns to be learnt have interrelationships. The desirability of such a scheme becomes evident when one considers say the requirement of identifying one's pet cat, in a collection of pictures dealing with other cats, besides those of the tiger, cheeta, leopard etc.

The model of Parga and Virasoro borrows heavily from spin glass physics. While they consider the existence of hierarchical relationship between the patterns stored, they do not, unlike Gidas, store and retrieve in a layer by layer fashion. Instead they merely renormalize or dress the J_{ij}'s of the Hebb learning rule. The manner in which this is accomplished is best discussed by considering a specific example.

Consider figure 49 where the "leaves" at the bottom denote the patterns to be learnt. The tree structure indicates the hierarchical relationship among the patterns in a notation that is self explanatory. Barring "Adam", there are three levels. The self overlap of a leaf with itself is denoted by q_3; it will be equal to 1. The overlap between two leaves derived from the same ancestor (in level 2) is denoted q_2; q_1 is defined similarly.

Given are the patterns corresponding to the leaves. Those corresponding to the ancestors i.e. patterns ζ^{11}, ζ^{12} etc. must be constructed from the information provided, and for this a rule is needed. Parga and Virasoro propose that if ζ_i^{11} is the spin of ζ^{11} at site i, then

$$\zeta_i^{11}=(\zeta_i^{111}+\zeta_i^{112})/2. \tag{98}$$

Similarly,

$$\zeta_i^1 = (\zeta_i^{11} + \zeta_i^{12})/2 \tag{99}$$

and so on. These expressions can be generalized in a straightforward manner.

In the usual Hebb learning rule, we have

$$J_{ij} = \sum_{\mu} \zeta_i^{\mu} \zeta_j^{\mu}$$

which would correspond to considering only the patterns in level 3. Parga and Virasoro modify this rule, besides adding contributions from other levels. For the hierarchy of figure 49 one would have, according to them,

$$NJ_{ij} = J_{ij}^{(3)} + J_{ij}^{(2)} + J_{ij}^{(1)}. \tag{100}$$

Here,

$$J_{ij}^{(1)} = \sum_{\alpha} \zeta_i^{\alpha} \zeta_j^{\alpha}/q_1, \quad \alpha = 1,2,3. \tag{101}$$

$$J_{ij}^{(2)} = \{1/(q_2 - q_1)\} \cdot \sum_{\alpha,\beta} (\zeta_i^{\alpha\beta} - \zeta_i^{\alpha})(\zeta_j^{\alpha\beta} - \zeta_j^{\alpha}), \; \alpha = 1,2,3 \;\; \beta = 1,2 \tag{102}$$

$$J_{ij}^{(3)} = \{1/(q_3 - q_2)\} \sum_{\alpha\beta\gamma} (\zeta_i^{\alpha\beta\gamma} - \zeta_i^{\alpha\beta})(\zeta_j^{\alpha\beta\gamma} - \zeta_j^{\alpha\beta}), \; \alpha = 1,2,3 \;\; \beta = 1,2 \;\; \gamma = 1,2 \tag{103}$$

Having learnt according to (100)–(102), retrieval is carried out as in the Hopfield model.

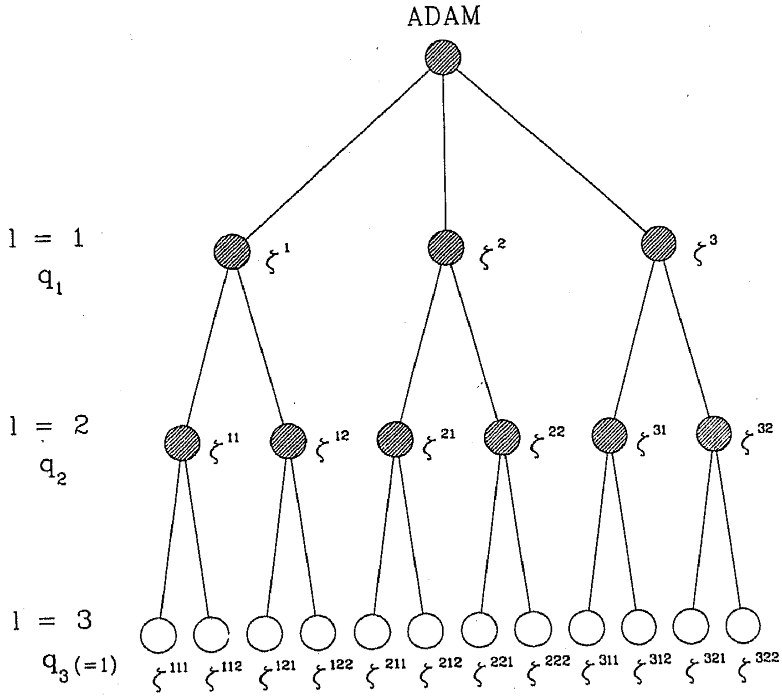


Figure 49. An elementary example of a set of patterns with a hierarchical relationship. There are three levels of relationship, the patterns to be learnt being at the level $l = 3$. All patterns are deemed to be derived from "Adam". q_1, q_2, q_3 are overlaps whose meanings are explained in the text.

Detailed statistical mechanical results for this model are not available, nor are numerical results. However, Parga and Virasoro have noted that their system can distinguish between spurious solutions belonging to well-defined categories of patterns and spurious solutions that mix categories. The former one is not serious but the latter one could be. As Parga and Virasoro remark, in the case of living beings, distinguishing between prey and predators is more vital than distinguishing among varieties of predators.

We now present a few illustrations of some of the above results. These are samples of an extensive study in progress, to benchmark the different learning algorithms. The detailed results of this study will be reported elsewhere. Most formal analysis, especially of the Hopfield model, concentrate on random input patterns whereas one would be more interested in storing and recalling non-random patterns. For this purpose a method was derived to generate a set of patterns which had some relationship among themselves.

The patterns were created as follows. To start with, the TSP was set up on a 8 × 8 grid, and locally optimal tours were generated for different, randomly chosen starting tours. These tours are then embedded into a 32 × 32 grid, whereupon the closed contour corresponding to the tour now encloses many more grid points. The interior points alone are now retained and the exterior points wiped out, giving a pattern. One example of how such a pattern is generated is shown in figures 50(a)–(c). Figure 50(d) shows a sample of a test patterns. In all, 1024 such patterns have been generated for our experiments. As can be seen, there is considerable overlap among the patterns.

The performance of the Hopfield model (i.e. with the Hebb/Cooper prescription) is illustrated in figure 51. Here, all except one pattern stored were random, the exception being one generated as described above. Results of recall with storage less than and exceeding critical capacity α_c are shown in the figure.

Of maximum interest perhaps are the results shown in figure 52. Unlike in the previous example, here all the patterns stored were generated via TSP tours, and therefore were presumably highly correlated. Samples of recalls based on learning via different rules are shown in the figure. As can be seen, the Hopfield model fares rather poorly.

7.6 *Beyond Hamiltonian models*

One feature common to all the models discussed so far is that in all of them,

$$J_{ij} = J_{ji}. \tag{104}$$

The merit of this symmetry is that it permits a Hamiltonian to be constructed. There are, however, strong reasons to believe that Nature favours neuronal asymmetry. It is in this context that models which relax the requirement (104) are considered. Of interest here is one due to Little (1975). This also considers a lattice of Ising spins whose spin state $\{S_i(t)\}$ is monitored at times $t = 0, \tau, 2\tau, \ldots$ etc. i.e. in time steps of length τ. The spin states evolve according to the equation

$$\rho\{N, (n+1)\tau\} = \sum_m W\{N, (n+1)\tau | M, n\tau\} \rho\{M, n\tau\} \tag{105}$$

Here M and N are particular spin states, with $\{M, n\tau\}$ denoting that M pertains to

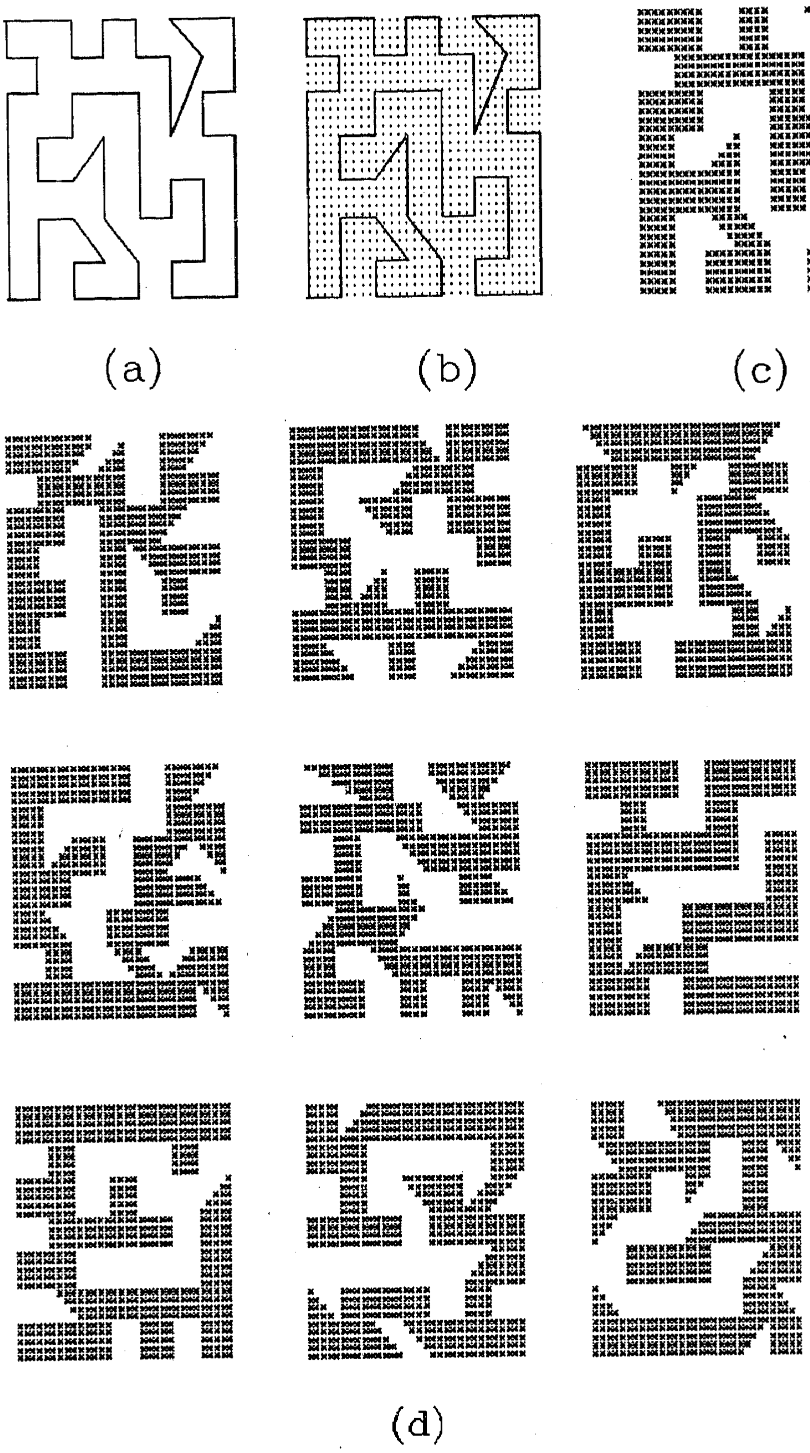


Figure 50. (a) shows a locally-optimal travelling salesman tour over a square lattice of size 8 × 8. This tour is now embedded into a square lattice of 32 × 32 as shown in (b). By retaining only the "interior" points, a pattern is created as in (c). A sample of nine such patterns is shown in (d).

the time $t = n\tau$. A similar meaning applies to $\{N, (n+1)\tau\}$. $\rho\{M, n\tau\}$ is the probability of finding the system in the state M at time $n\tau$, and $W\{N, (n+1)\tau | M, n\tau\}$ is the probability for the system to go to the state N at time $(n+1)\tau$, given that it was in

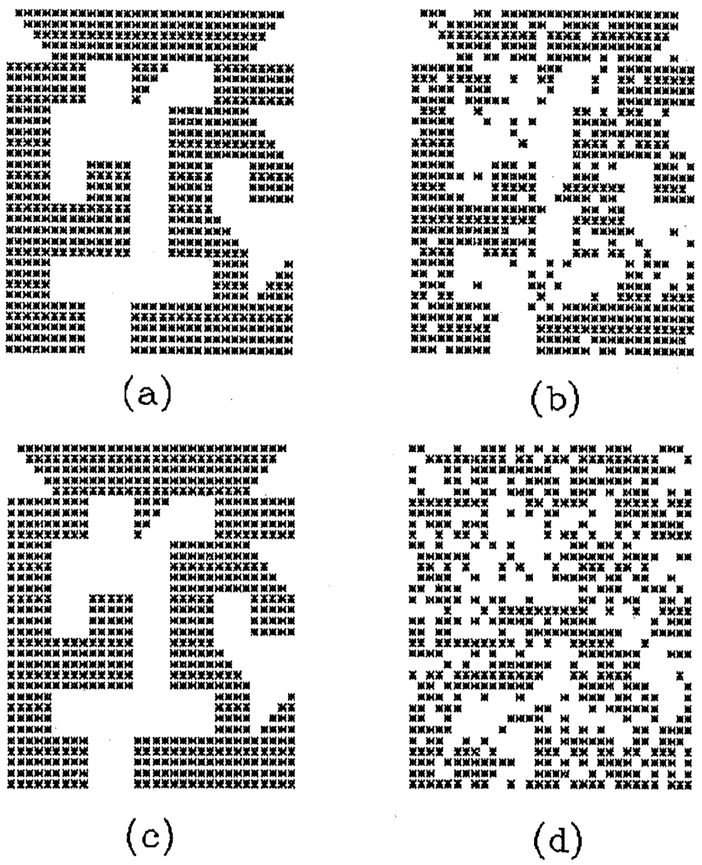


Figure 51. Results of two recall-experiments, corresponding to two different storage capacities. The patterns are taught to the network using the Hebb/Cooper learning rule. Among the stored patterns, one is prepared as described in the previous figure while the rest are random patterns. (a) shows the stored (non random) pattern. (b) is the noisy, input test pattern. It has 25% noise. (c) and (d) show the recalls corresponding to $\alpha = 0{\cdot}1$ $(< \alpha_c)$ and $\alpha = 0{\cdot}25$ $(< \alpha_c)$.

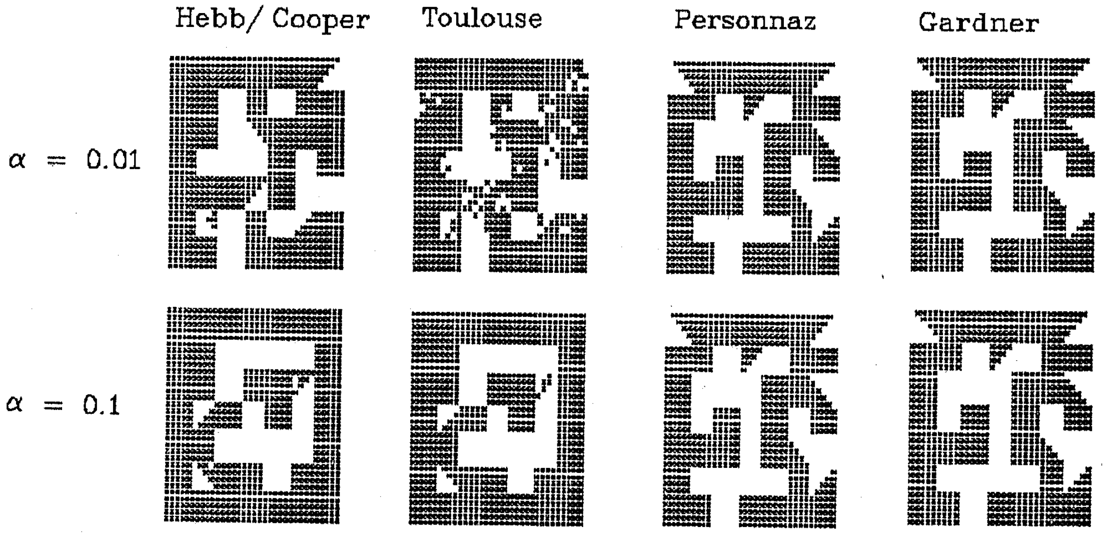


Figure 52. Results of recall as obtained with Hebb/Cooper, Toulouse, Personnaz and Gardner learning rules. Results are shown for $\alpha = 0{\cdot}01$ and $\alpha = 0{\cdot}1$. All the stored patterns were generated as described in figure 51. They therefore had considerable overlap. The input test pattern is the same as in figure 51 (b).

M at time $n\tau$. The W's obey the stationarity condition

$$W\{N,(n+1)\tau|M,n\tau\} = W\{N,\tau|M,0\} \tag{106}$$

and the normalization constraint

$$\sum_N W\{N,\tau|M,0\} = 1. \tag{107}$$

The transition probability $\omega(N|M)$ per unit time for the transition $M \rightarrow$ N is defined by

$$\omega(N|M) = \frac{W\{N,\tau|M,0\}}{\tau}, \tag{108}$$

and in the Little model it is given by

$$\omega(N|M) = \exp[-\beta\mathscr{H}(N|M)] \Big/ \sum_P \exp[-\beta\mathscr{H}(P|M)] \tag{109}$$

where

$$\mathscr{H}(N|M) = -\sum_{ij} J_{ij} S_i(N) S_j(M). \tag{110}$$

The meaning of $S_i(N)$ and $S_i(M)$ should be clear. Observe that S_i and S_j refer to *different times.*

So far, no Hamiltonian has entered the picture although $\mathscr{H}(N/M)$ might look like one. If $J_{ij} = J_{ji}$, the stationary states of the Little model can indeed be derived from a Hamiltonian (Peretto 1984) which is of the form

$$\mathscr{H}_{\text{Little}} = -(1/\beta)\sum_i \ln\left[2\cosh\left(\beta\sum_j J_{ij}S_j\right)\right]. \tag{111}$$

Amit *et al* (1985) have shown that the thermodynamic properties and the long-term behaviour of the Little model (as described by (111)) are the same as those of the (discrete) Hopfield model.

Passing reference may be made here about the relative merits of the digital and the analog versions of neural networks. Incidentally, Hopfield and Tank have argued strongly in favour of the latter. Getting back to the question of a comparison, we have noted that in discrete models there could be a trapping in a spurious valley. What happens in the case of analog networks? Here also the equations of motion drive the system to lower energy states. However, the use of "continuous" neurons smoothens the energy landscape and there are no shallow wells. The system can therefore roll down to the bottom of a deep valley, with as much facility as can be provided by simulated annealing to a discretized version of the system. Such a solution (rolling down a smoothened slope) may not be optimal but "good" and therefore acceptable. Indeed, biological systems may well be attuned to selecting "good" solutions rather than optimal ones.

7.7 *Neural nets – a parting overview*

The subject of neural networks is vast and what we have reviewed is restricted to areas where physicists have been active lately. Here it is pertinent to cite Grossberg

(1988) who gives an absorbing summary of how, over the years, physics and physiology became dissociated although giants like Helmholtz and Maxwell once studied both. Much later, mathematics and physics began to penetrate physiology once again, largely through models for various physiological processes. Perhaps the most famous of these is the Hodgkin-Huxley model for nerve conduction. The point sought to be stressed in the present paper is that experience in many-body theory and statistical mechanics now appears to be quite relevant in the study of neural networks. One anticipates that in the years to come, there would be inputs also from analytical dynamics where large expertise has been accumulated concerning attractors. In short, exciting and new vistas are opening up for physicists, wherein their knowledge could come in handy (see, for example, Haken 1978, 1988) to achieve breakthroughs in a frontier area like neural networks. In this context, it is interesting to note that the European Economic Community has launched a multi-million dollar programme called BRAIN (basic research in adaptive research and neuro computing), to promote cooperation among biologists, computer scientists and physicists interested in neural networks.

Acknowledgements

We would like to thank Dr M V Atre for useful discussions and Mr A Ratnakar for invaluable help in literature survey. We are also grateful to the American Physical Society, Journal de Physique and the Journal of Statistical Physics for permission to reproduce figures earlier published by them. One of us (GV) would like to thank Dr V K Viswanathan for drawing our attention to many pertinent references, besides providing reprints of some of them. Finally, we are indebted to Mrs D Manikyamma and Mrs Naga Nirmala for assistance with the preparation of the manuscript.

References

Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* **A32** 1007
Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30
Anderson P W 1978 *Rev. Mod. Phys.* **50** 199
Barahona F, Maynard R, Rammal R and Uhry J P 1982 *J. Phys.* **A15** 673
Beardwood J, Halton J H and Hammersley J M 1959 *Proc. Cambridge Philos. Soc.* **55** 299
Bieche I, Maynard R, Rammal R and Uhry J P 1980 *J. Phys.* **A53** 2553
Binder K and Young A P 1986 *Rev. Mod. Phys.* **58** 801
Blandin A, Gabay M and Garel T 1980 *J. Phys.* **C13** 403
Bohachevsky I O, Johnson M E and Stein M L 1986 *Technometrics* **28** 209
Bohachevsky I O, Johnson M E and Stein M L 1987 preprint of a paper submitted to *Naval Research Logistics Quarterly Journal.*
Bonomi E and Lutton J L 1984 *SIAM Rev.* **26** 551
Bounds D G 1987 *Nature (London)* **329** 215
Brady R M 1985 *Nature (London)* **317** 804
Bray A J and Moore M A 1980 *J. Phys.* **C13** L469
Bray A J, Moore M A and Young A P 1984 *J. Phys.* **C17** L-155
Cahn R W 1980 *Contemp. Phys.* **21** 43
Cerny V 1985 *J. Optim. Theory Appl.* **45** 41
Cooper L N 1973 *Nobel Symp.* **24** 252

Cox A D and Youngman A P 1985 *Proceedings of the International Conference on Insertion Devices for Synchrotron Sources* (Washington: Society of Photo-Optical Instrumentation Engineers) SPIE **582** 91

de Almeida J R L and Thouless D J 1978 *J. Phys.* **A11** 5

de Dominicis C, Gabay M, Garel T and Orland H 1980 *J. Phys.* (*Paris*) **41** 923

Dotsenko V S 1985 *J. Phys.* **C18** L1017

Edmonds J 1965 *Oper. Res.* **13** Suppl. 1 373

Edmonds J and Johnson E L 1973 *Math. Prog.* **5** 88

Edwards S F and Anderson P W 1975 *J. Phys.* **F5** 965

Ford P J 1982 *Contemp. Phys.* **23** 141

Gardner E 1988 *J. Phys.* **A21** 257

Garey M R and Johnson D S 1979 *Computers and Intractability* (San Francisco: Freeman)

Gidas B 1989 *IEEE Trans. Patt. Anal and Mach. Int.* **11** 164

Glauber R J 1963 *J. Math. Phys.* **4** 294

Graf H P, Tackel L D, Howard R E, Staughn B, Denker J S, Hubbard W, Tennant D M and Schwartz D 1986 *Neural Networks for Computing AIP. Conf. Proc.* **151** 182

Grossberg S 1988 *Neural Networks* **1** 17

Haken H 1978 *Synergetics: An Introduction* (Berlin: Springer)

Haken H 1988 *Z. Phys.* **B70** 121

Hebb D O 1949 *The Organisation of Behaviour* (New York: Wiley)

Hopfield J J 1982 *Proc. Natl. Acad: Sci. USA* **79** 2554

Hopfield J J 1984 *Proc. Natl. Acad. Sci. USA* **81** 3088

Hopfield H J J and Tank D W 1985 *Biol. Cybern.* **52** 141

Hopfield J J and Tank D W 1986 *Science* **233** 625

Kanter I and Sompolinsky 1987 *Phys. Rev.* **A35** 380

Kimura M 1985 *New Scientist* July 11 41

Kirkpatrick S 1981 *Models of Disordered Systems Lecture Notes in Physics* **149** (Berlin: Springer) p. 280

Kirkpatrick S 1984 *J. Stat. Phys.* **34** 975

Kirkpatrick S, Gelatt C D (Jr) and Vecchi M P 1983 *Science* **220** 671

Kirkpatrick S and Sherrington D 1978 *Phys. Rev.* **B17** 11

Kirkpatrick S and Toulouse G 1985 *J. Phys.* **46** 1277

Lin S 1965 *Bell Syst. Tech. J.* **44** 2245

Lin S and Kernighan B W 1973 *Oper. Res* **21** 498

Lippmann R P 1978 *IEEE ASSP Mag.* April Issue 4

Little W A 1975 *Math. Biosci.* **19** 101

Ma S K 1985 *Statistical Mechanics* (Singapore: World Scientific)

Mei-Ko Kwan 1962 *Chinese Math.* **1** 273

Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087

Mezard M and Parisi G 1985 *J. Phys.* (*Paris*) **46** L771

Mezard M, Parisi G, Sourlas N, Toulouse G and Virasoro M 1984a *Phys. Rev. Lett.* **52** 13

Mezard M, Parisi G, Sourlas N, Toulouse G and Virasoro M 1984b *J. Phys.* (*Paris*) **45** 843

Mezard M, Parisi G and Virasoro M 1986 *Europhys. Lett.* **1** 77

Orland H 1985 *J. Phys.* (*Paris*) *Lett.* **46** L-763

Palmer R G 1982 *Adv. Phys.* **31** 669

Palmer R G 1983 *Heidelberg Colloquium on Spin Glasses*, Lecture Notes on Physics 192 (ed.) J L van Hemmen and I Morgenstern (Berlin: Springer) 234

Parga N and Virasoro M A 1986 *J. Phys.* (Paris) **47** 1857

Parisi G 1979 *Phys. Rev. Lett* **43** 23

Parisi G 1980a *J. Phys.* **A13** 1101

Parisi G 1980b *J. Phys.* **A13** 1887

Parisi G 1980c *J. Phys.* **A13** L115

Parisi G 1983 *Phys. Rev. Lett.* **50** 24

Parisi G 1986 *J. Phys.* **A19** L675

Peretto P 1984 *Biol. Cybern.* **50** 24

Personnaz L, Guyon I and Dreyfus G 1985 *J. Phys.* (*Paris*) *Lett.* **46** L359

Rammal R, Toulouse G and Virasoro M 1986 *Rev. Mod. Phys.* **58** 765

Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 26

Siarry P and Dreyfus G 1984 *J. Phys. (Paris) Lett.* **45** L-39

Sivilotti M A, Emerling M R and Mead C A 1986 *Neural Networks for Computing AIP Conf. Proc.* **151** 408

Sompolinsky H 1988 *Phys. Today* December 70

Southern B W 1987 *Can. J. Phys.* **65** 1245

Stanley H E 1971 *An Introduction to Phase Transitions and Critical Phenomena* (Oxford: University Press)

Tank D W and Hopfield J J 1987 *Sci. American* December 62

Tanaka F and Edwards S F 1980 *J. Phys.* **F10** 2471

Thouless D J, Anderson P W and Palmer R G 1977 *Philos. Mag.* **35** 593

Toulouse G 1983 *Heidelberg Colloquium on Spin Glasses* Lecture Notes in Physics 192 (ed.) J L van Hemmen and I Morgenstern (Berlin: Springer) 2

Toulouse G, Dehaene S and Changeux J 1986 *Proc. Natl. Acad. Sci. USA* **83** 1695

van Hemmen J L 1983 *Heidelberg Colloquium on Spin Glasses Lecture Notes in Physics* 192 (ed.) J L van Hemmen and I Morgenstern (Berlin: Springer) 203

van Hemmen J L and Palmer R G 1979 *J. Phys.* **A12** 563

Vannimenus J and Mezard M 1984 *J. Phys. (Paris) Lett.* **45** L-1145

Viswanathan V K, Bohachevsky I O and Cotter T P 1985 *Proceedings of the International Lens Design Conference* (Washington: Society of Photo-Optical Instrumentation Engineers) SPIE 554 10

Williams G 1987 *Can. J. Phys.* **65** 1251

Young A P 1983 *Phys. Rev. Lett.* **51** 1206

Young A P, Bray A J and Moore M A 1984 *J. Phys.* **C17** L149