Optimal storage capacity of neural networks at finite temperatures

1993 J. Phys. A: Math. Gen. 26 3741

(http://iopscience.iop.org/0305-4470/26/15/024)

J. Phys. A: Math. Gen. 26 (1993) 3741–3755. Printed in the UK

Optimal storage capacity of neural networks at finite temperatures

G M Shim†, D Kim‡ and M Y Choi‡

Department of Physics and Centre for Theoretical Physics, Seoul National University, Seoul 151-742, Korea

Received 3 November 1992, in final form 4 May 1993

Abstract. Gardner's analysis of the optimal storage capacity of neural networks is extended to study finite-temperature effects. The typical volume of the space of interactions is calculated for strongly diluted networks as a function of the storage ratio α, temperature T and the tolerance parameter m, from which the optimal storage capacity α_c is obtained as a function of T and m. At zero temperature it is found that $\alpha_c = 2$ regardless of m while α_c in general increases with the tolerance at finite temperatures. We show how the best performance for given α and T is obtained, which reveals a first-order transition from high-quality performance to a low-quality one at low temperatures. An approximate criterion for recalling, which is valid near $m = 1$, is also discussed.

1. Introduction

Recently, the tools of statistical mechanics have been extensively applied to the study of collective properties of neural networks [1]; spin glass theory has played an important role in the growth of this new field [12]. In particular, the optimal (error-free) storage capacity for recurrent networks can be obtained by calculating the typical fractional volume of the space of interactions ($\{J_{ij}\}$) satisfying the condition that, for a given set of patterns, each pattern is a fixed point of the deterministic (zero-temperature) dynamics

$$s_i(t+1) = \operatorname{sign}\left[\sum_j J_{ij} s_j(t)\right] \tag{1}$$

where $s_i(t)$ $(= \pm 1)$ $(i = 1, \ldots, N)$ represents the state of the ith neuron at time t, and the synaptic coupling J_{ij} determines the contribution of a signal fired by the jth neuron to the action potential on the ith neuron. This approach to systematic exploration of the space of interactions, which was pioneered by Gardner [3] and reformulated in terms of canonical ensemble calculation [4], has been applied in various directions [4–11]. The Hopfield model with general continuous couplings has been found to be capable of storing at most two uncorrelated random patterns per neuron without errors and larger number of patterns for biased patterns [3]. The network with discrete (Ising-type) couplings has also been extensively investigated since the replica-symmetry theory was reported to yield incorrect results for the optimal storage capacity [4, 6, 7]. The method is not limited to

† Present address: Institute voor Theoretische Fysica, Katholieke Universiteit Leuven, B–3001 Leuven, Belgium. e-mail: fgbda28@cc1.kuleuven.ac.be

‡ e-mails: dkim@phya.snu.ac.kr, mychoi@phya.snu.ac.kr

0305-4470/93/153741+15$07.50

Hopfield-type neural networks but applicable to multilayer networks as well as to simple perceptrons [5, 11]. However, Gardner's method is based on the concept of fixed points of the dynamics and, clearly, does not work if the dynamics is stochastic (i.e. at finite temperatures). In addition, it requires perfect matching so that each pattern is unerringly recalled at every site whereas, in practice, one usually considers a neural network to 'remembering' or 'recalling' if the overlap between the network state and one of the patterns is larger than some given value.

In this paper, we propose a scheme to define the optimal storage capacity at finite temperatures and study its temperature dependence. We introduce the tolerance parameter $m(\leqslant 1)$ in such a way that the $m \to 1$ limit corresponds to the perfect recall while $(1-m)/2$ measures the error allowed. We then calculate the typical fractional volume of the space of interactions for extremely diluted networks as a function of the storage ratio α, temperature T and the tolerance parameter m, which leads to the optimal storage capacity α_c as a function of T and m. At zero temperature it is found that $\alpha_c = 2$ regardless of the tolerance parameter m. At finite temperatures, on the other hand, the optimal storage capacity vanishes in the perfect matching limit ($m \to 1$) and, in general, increases with the tolerance. We then discuss how the best performance is obtained for given α and T. We also propose an alternative criterion for recalling, which may be regarded as a simple approximate scheme to define the optimal storage capacity, and consider the optimal storage capacity of the dynamic model [12] as well as of the extremely diluted model in this approximate scheme.

The contents of this paper are as follows. In section 2, we propose a scheme to define the optimal storage capacity at finite temperatures together with an approximate scheme. Section 3 is devoted to the calculation of the optimal storage capacity of an extremely diluted neural network while section 4 presents results of the approximate scheme. A brief discussion is given in section 5.

2. Optimal storage capacity at finite temperatures

One usually takes internal noise in the functioning of a neuron into account by extending the deterministic evolution rule (1) to a stochastic one:

$$P[s_i(t+1) = \pm 1] = \frac{1}{2}\left\{1 \pm \tanh\left[\beta \sum_j J_{ij} s_j(t)\right]\right\} \tag{2}$$

where the inverse temperature ($\beta \equiv 1/T$) measures the width of the threshold region, i.e. the level of synaptic noise. The state $s \equiv \{s_i\}$ of the network of N neurons evolves stochastically according to equation (2). A given set of states of the network to be memorized by appropriately adjusting the couplings is called the set of patterns. We now define the overlap $M_\mu(t)$ between the network state and the μth pattern $\xi^\mu \equiv \{\xi_i^\mu\} (\mu = 1, \ldots, p)$ by

$$M_\mu(t) \equiv \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu s_i(t)$$

which also evolves stochastically along with $s(t)$. When a network is recalling pattern μ, the time average of $M_\mu(t)$ over a time scale sufficiently longer than the observational time but shorter than the lifetime of the local energy minimum should be close to unity. However, since the dynamics is stochastic, it cannot be strictly unity as in zero-temperature

dynamics. Therefore, we introduce the tolerance parameter m in such a way that the network is considered to be remembering the μth pattern if

$$\overline{M}_\mu \equiv \frac{1}{N_t}\sum_{t=1}^{N_t} M_\mu(t) > m$$

with N_t in the appropriate range as mentioned above. The quantity $(1-m)/2$ is the maximum error allowed for the network to be qualified as *functioning*. It is expected that the time average $\overline{M}_\mu$ will be equivalent to the restricted thermal average

$$\langle M_\mu \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \langle s_i \rangle$$

where the thermal measure is restricted within a single pure state (containing the configuration ξ^μ) [13]. In the stationary state the activity $\langle s_i \rangle$ of the ith neuron is determined by the coupled equations

$$\langle s_i \rangle = \left\langle \tanh\left(\beta \sum_j J_{ij} s_j\right)\right\rangle = \tanh\left(\beta \sum_j J_{ij}\langle s_j \rangle\right)$$

where a mean-field approximation has been used. Such an approximation is expected to be valid for diluted networks which we mainly consider in this work. Otherwise, a reaction term may be necessary. The optimal storage capacity is given by the upper bound of the *storage ratio* $\alpha \equiv p/N$, *where* p *is the number of stored patterns. The problem reduces*, according to Gardner [3], to the calculation of the typical fractional volume of the space of interactions which satisfies the following conditions:

$$\frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \tanh\left(\frac{\beta}{\sqrt{N}}\sum_j J_{ij}\langle s_j \rangle\right) > m \tag{3}$$

and

$$\sum_j (J_{ij})^2 = N \qquad \text{for each } i. \tag{4}$$

The condition in equation (4) is required to fix the scale of temperature T. The optimal storage capacity at temperature T is then determined by eliminating this fractional volume, which leads to α_c as a function of T and m. This scheme will be applied to an extremely diluted neural network in the following section. In this model, only an extremely small fraction of the couplings among neurons are connected so that its dynamics can be solved in a simple manner [14, 15]. However, calculation of the fractional volume for other generic neural networks is formidable as the thermal average has to be performed within one pure state. As a simple attempt, one may use the approximation $\langle s_i \rangle \approx \xi_i^\mu$ and replace equation (3) by

$$\frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \tanh\left(\frac{\beta}{\sqrt{N}}\sum_j J_{ij}\xi_j^\mu\right) > m \tag{5}$$

which states that the network state evolved by one time step from a given pattern has overlap with that pattern greater than m. This simple criterion presumably leads to results similar to those of equation (3) for m close to unity, where the network is expected to hover around the configuration ξ^μ during the recalling state. The validity of this approximate scheme with regard to diluted networks is discussed in section 4.

3. Extremely diluted neural network

In this section the proposed scheme is applied to an extremely diluted neural network, where, on average, there are $C(\lesssim \log N)$ connections per neuron. Such a model was first studied by Derrida *et al* [14], and its properties of the basin of attraction was studied later by Gardner [15] and Amit *et al* [16]. The reason we can implement the scheme exactly is that the dynamics of the network can be solved in a simple manner [14, 15]. Extending the method of Keppler and Abbott [17], we describe the time evolution of the overlap $M_\mu(t)$ between pattern ξ^μ and the network state by the one-step recursion relation

$$M_\mu(t+1) = F_{h^\mu}[M_\mu(t)]. \tag{6}$$

Here the *map* $F_{h^\mu}(x)$ is defined by

$$F_{h^\mu}(x) \equiv \frac{1}{N}\sum_{i=1}^{N}\int \mathrm{D}z \tanh\left[\beta\left(\sqrt{1-x^2}z + h_i^\mu x\right)\right]$$

where $\beta(\equiv T^{-1})$ is the inverse temperature, $\mathrm{D}z$ denotes the Gaussian measure: $\mathrm{D}z \equiv \exp(-z^2/2)\,\mathrm{d}z/\sqrt{2\pi}$, and the subscript h^μ denotes the $\{h_i^\mu\}$-dependence of the map with

$$h_i^\mu \equiv \xi_i^\mu \sum_j \frac{J_{ij}}{\sqrt{C}}\xi_j^\mu. \tag{7}$$

In the stationary state, $M_\mu(t)$ approaches $M_\mu(t\to\infty) = M_\mu^*$ the value of which is determined by the stable fixed-point solution x satisfying

$$x = F_{h^\mu}(x) \tag{8a}$$

$$\left|\frac{\partial}{\partial x}F_{h^\mu}(x)\right| \equiv |F'_{h^\mu}(x)| < 1 \tag{8b}$$

where equation (8*b*) has been imposed to guarantee its stability. For given tolerance parameter m, the network is considered to be remembering the μth pattern if the value of the stationary overlap M_μ^* is greater than m.

Now the main quantity to calculate is the fractional volume of the space of interactions ($\{J_{ij}\}$) for which every pattern can be remembered. The normalization condition is now given by

$$\sum_j (J_{ij})^2 = C$$

instead of equation (4). The number of the solutions of equation (8) with its value greater than m is formally given by

$$\mathcal{N}_{h^\mu} \equiv \int_m^1 \mathrm{d}M\, \delta(M - F_{h^\mu}(M))|1 - F'_{h^\mu}(M)|\theta(1 - |F'_{h^\mu}(M)|) \tag{9}$$

so that the fractional volume can be written as

$$V_0 = \frac{\int\left[\prod_{i\neq j}\mathrm{d}J_{ij}\right]\prod_i \delta(\sum_j (J_{ij})^2 - C)\prod_{\mu=1}^{\alpha C}\theta(\mathcal{N}_{h^\mu})}{\int\left[\prod_{i\neq j}\mathrm{d}J_{ij}\right]\prod_i \delta(\sum_j (J_{ij})^2 - C)}$$

where $\theta(x)$ is the step function and the number of stored patterns has been scaled according to $p \equiv \alpha C$. Here we are mainly interested in obtaining the optimal storage capacity α_c. If the number of stored patterns exceeds $\alpha_c C$, there is no typical network ($\{J_{ij}\}$) that yields the value of the stationary overlap which is greater than m for all patterns. In the limit $\alpha \to \alpha_c$, the number $\mathcal{N}_{h^\mu}$ of stable solutions approaches zero, and we may replace the step function $\theta(\mathcal{N}_{h^\mu})$ by $\mathcal{N}_{h^\mu}$. Furthermore the fractional volume vanishes only if $\mathcal{N}_{h^\mu} = 0$ for some μ, which implies that the replacement $\theta(x) \to x$ would not affect the optimal storage capacity. The fractional volume to calculate is now given by

$$V = \frac{\int \left[\prod_{i \neq j} \mathrm{d}J_{ij}\right] \prod_i \delta(\sum_j (J_{ij})^2 - C) \prod_\mu \mathcal{N}_{h^\mu}}{\int \left[\prod_{i \neq j} \mathrm{d}J_{ij}\right] \prod_i \delta(\sum_j (J_{ij})^2 - C)}. \tag{10}$$

Replacement of $\theta(x)$ by x is, in general, also justified in the following sense. We may assume that, for typical $\{\xi^\mu\}$, $\mathcal{N}_{h^\mu}$ possesses a finite system-size-independent upper bound almost everywhere in the interaction space. This is reasonable since the map $F_{h^\mu}(x)$ is an average of N functions of the form

$$\int \mathrm{D}z \tanh\left[\beta\left(\sqrt{1-x^2}z + hx\right)\right].$$

This function is smooth and monotonic in x with a derivative having the same sign as h. Therefore the average over many possible h is a sum of two parts: the monotonically increasing part from contributions of $h > 0$ and the monotonically decreasing part from $h < 0$. So, in practice, there are only a few solutions at most. If we denote the upper bound by $\mathcal{N}_0$, we then have the identity

$$\theta(\mathcal{N}_{h^\mu}) \leqslant \mathcal{N}_{h^\mu} \leqslant \mathcal{N}_0 \theta(\mathcal{N}_{h^\mu}).$$

Integrating this over the interaction space, we immediately see that

$$V_0 \leqslant V \leqslant \mathcal{N}_0^{\alpha C} V_0.$$

However, the fractional volumes are of the order of $\exp(-CN)$ so that $(\log V)/CN$ is the same as $(\log V_0)/CN$ in the thermodynamic limit $N \to \infty$.

In this work, we consider the case that every pattern ξ_i^μ to be stored is an independently distributed random variable, taking the value ± 1 with equal probabilities. The typical fractional volume $\overline{V} \equiv \exp(\langle(\log V)\rangle)$ for the random patterns involves averaging $\log V$ over the distribution of the random patterns $\{\xi^\mu\}$, which may be obtained through the use of the well known replica trick. To facilitate the averaging over the distribution of the random patterns, we introduce δ-functions describing equation (7) with the help of the conjugate variable $\hat{h}_i^\mu$ raised to the exponential form

$$\delta\left(h_i^\mu - \xi_i^\mu \sum_j \frac{J_{ij}}{\sqrt{C}} \xi_j^\mu\right) = \int \frac{\mathrm{d}\hat{h}_i^\mu}{2\pi} \exp\left[\mathrm{i}\hat{h}_i^\mu \left(h_i^\mu - \xi_i^\mu \sum_j \frac{J_{ij}}{\sqrt{C}} \xi_j^\mu\right)\right].$$

The average over the random patterns for the replicated volume $\langle\langle V^n \rangle\rangle$ affects the exponential factor containing ξ_i^μ in the above expression, and leads to the following:

$$\left\langle\!\!\left\langle \prod_{\alpha=1}^{n} \exp\left(-\mathrm{i} \sum_{\mu i} \hat{h}_i^{\mu\alpha} \xi_i^\mu \sum_j \frac{J_{ij}^\alpha}{\sqrt{C}} \xi_j^\mu \right) \right\rangle\!\!\right\rangle_{\{\xi^\mu\}}$$

$$= \prod_\mu \exp\left(-\tfrac{1}{2} \sum_{\alpha\beta i} \hat{h}_i^{\mu\alpha} \hat{h}_i^{\mu\beta} \sum_j \frac{J_{ij}^\alpha J_{ij}^\beta}{C} - \tfrac{1}{2} \sum_{\alpha i} \hat{h}_i^{\mu\alpha} \sum_{\beta j} \hat{h}_j^{\mu\beta} \frac{J_{ij}^\alpha J_{ji}^\beta}{C} \right)$$

where α and β are the replica indices and, for an extremely diluted network, the cumulant expansion has been cut-off at the second order [8]. Following [8], we assume the second term to be independent of site i, so that

$$\sum_{\beta j} \hat{h}_j^{\mu\beta} \frac{J_{ij}^{\alpha} J_{ji}^{\beta}}{C} = \frac{1}{N} \sum_i \sum_{\beta j} \hat{h}_j^{\mu\beta} \frac{J_{ji}^{\beta} J_{ij}^{\alpha}}{C}.$$

Introducing the local order parameters

$$q_{\alpha\beta}^i \equiv \frac{1}{C} \sum_j J_{ij}^{\alpha} J_{ij}^{\beta} \qquad r_{\alpha\beta}^i \equiv \frac{1}{C} \sum_j J_{ij}^{\alpha} J_{ji}^{\beta}$$

together with their respective conjugate variables $Q_{\alpha\beta}^i$ and $R_{\alpha\beta}^i$, we obtain

$$\langle\langle V^n \rangle\rangle = \frac{1}{\exp(nCN/2)} \int \prod_{\alpha i} \frac{\mathrm{d}E_i^{\alpha}}{4\pi \mathrm{i}} \int \prod_{\alpha<\beta\, i} \frac{\mathrm{d}Q_{\alpha\beta}^i \mathrm{d}q_{\alpha\beta}^i}{2\pi \mathrm{i}/C} \int \prod_{\alpha\beta i} \frac{\mathrm{d}R_{\alpha\beta}^i \mathrm{d}r_{\alpha\beta}^i}{4\pi \mathrm{i}/C} \exp(CG)$$

where

$$G \equiv \tfrac{1}{2} \sum_{\alpha i} E_i^{\alpha} + \sum_{\alpha<\beta\, i} Q_{\alpha\beta}^i q_{\alpha\beta}^i + \tfrac{1}{2} \sum_{\alpha\beta i} R_{\alpha\beta}^i r_{\alpha\beta}^i + G_1 + \alpha G_2$$

with G_1 and G_2 given by

$$\exp(CG_1) \equiv \int \left[\prod_{\alpha\, i\neq j} \mathrm{d}J_{ij}^{\alpha} \right] \exp\left(-\tfrac{1}{2} \sum_{\alpha i j} E_i^{\alpha} (J_{ij}^{\alpha})^2 - \sum_{\alpha<\beta} Q_{\alpha\beta}^i J_{ij}^{\alpha} J_{ij}^{\beta} - \tfrac{1}{2} \sum_{\alpha\beta i j} R_{\alpha\beta}^i J_{ij}^{\alpha} J_{ji}^{\beta} \right)$$

$$\exp(G_2) \equiv \int \left[\prod_{\alpha i} \frac{\mathrm{d}\hat{h}_i^{\alpha} \mathrm{d}h_i^{\alpha}}{2\pi} \right] \left[\prod_{\alpha} \mathcal{N}_{h^{\alpha}} \right] \exp\left(\mathrm{i} \sum_{\alpha i} \hat{h}_i^{\alpha} h_i^{\alpha} - \tfrac{1}{2} \sum_{\alpha<\beta\, i} \hat{h}_i^{\alpha} \hat{h}_i^{\beta} q_{\alpha\beta}^i \right.$$
$$\left. - \frac{1}{2N} \sum_{\alpha\beta i j} \hat{h}_i^{\alpha} \hat{h}_j^{\beta} r_{\alpha\beta}^j - \tfrac{1}{2} \sum_{\alpha i} (\hat{h}_i^{\alpha})^2 \right).$$

In the thermodynamic limit ($N \to \infty$), C also approaches infinity, albeit slowly, and $\langle\langle V^n \rangle\rangle$ can be computed through the use of the steepest-descent method. In order to find the saddle point, we assume the replica- and site-symmetric ansatz

$$E_i^{\alpha} = E \qquad R_{\alpha\alpha}^i = S \qquad r_{\alpha\alpha}^i = s$$
$$Q_{\alpha\beta}^i = Q \qquad q_{\alpha\beta}^i = q \qquad R_{\alpha\beta}^i = R \qquad r_{\alpha\beta}^i = r \qquad (\alpha \neq \beta).$$

With this ansatz, the function G in the limit $n \to 0$ takes the form

$$G = nN[\tfrac{1}{2}(E - qQ + sS - rR) + g_1 + \alpha g_2]$$

where

$$g_1 \equiv -\tfrac{1}{4} \left(\frac{Q+R}{E-Q+S-R} + \frac{Q-R}{E-Q-S+R} + \log(E-Q+S-R) + \log(E-Q-S+R) \right)$$

$$g_2 \equiv \frac{1}{N} \int \prod_{i=1}^{N} \mathrm{D}t_i \log \left[\int \frac{\mathrm{d}t_0}{\sqrt{2\pi/N}} \exp(-\tfrac{1}{2} N t_0^2) \int \prod_{i=1}^{N} \mathrm{D}h_i \mathcal{N}_H \right].$$

In the above expression $\mathcal{N}_{\boldsymbol{H}}$ is given by equation (9) with h^μ replaced by $\boldsymbol{H} \equiv \{H_i\}$, where $H_i \equiv \sqrt{1-q}\, h_i - \sqrt{s-r}\, t_0 - \sqrt{q}\, t_i$. Since the saddle-point equations for the variables E, S, Q and R are algebraic, we can eliminate these variables and finally write the typical fractional volume in the form

$$\overline{V} = \exp\left(CN\left[\frac{1}{2}\log(1-q) + \frac{1}{4}\log(1-x^2) + \frac{1}{2}\frac{q-rx}{(1-q)(1-x^2)} + \alpha g_2 \right] \right)$$

where $x \equiv (s-r)/(1-q)$.

To manipulate g_2, we introduce the variable

$$\overline{M} \equiv \frac{\partial}{\partial M} F_{\boldsymbol{H}}(M)$$

together with its conjugate variable $\overline{\lambda}$ and use the integral representation of δ-function. Noting the range of the variable $\overline{M}$, we obtain

$$g_2 = \frac{1}{N}\int_{-\infty}^{\infty} \prod_{i=1}^{N} \mathrm{D}t_i \log\left[\int_{-\infty}^{\infty} \frac{\mathrm{d}t_0}{\sqrt{2\pi/N}} \int_m^1 \mathrm{d}M \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}\lambda}{2\pi\mathrm{i}/N} \int_{-1}^{1} \mathrm{d}\overline{M} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}\overline{\lambda}}{2\pi\mathrm{i}/N} (1-\overline{M}) \right.$$
$$\left. \times \exp[-N(\tfrac{1}{2}t_0^2 + \lambda M + \overline{\lambda}\overline{M})] \prod_{i=1}^{N} \int \mathrm{D}h_i \exp[\lambda f(H_i, M) + \overline{\lambda}\overline{M})]\partial_M f(H_i, M)] \right]$$

where the functions $f(H, M)$ and $\partial_M f(H, M)$ are given by

$$f(H, M) \equiv \int \mathrm{D}z \tanh[\beta(\sqrt{1-M^2}\, z + MH)]$$
$$\partial_M f(H, M) \equiv \partial f(H, M)/\partial M.$$

Now the integration over $\overline{M}$ is easily performed and, in the thermodynamic limit, the steepest-descent method yields

$$g_2 = \max_{m \leqslant M \leqslant 1} \max_{t_0} \min_{\lambda, \overline{\lambda}} \left(-\tfrac{1}{2}t_0^2 - \lambda M + |\overline{\lambda}| \right.$$
$$\left. + \int \mathrm{D}t \log \int \mathrm{D}h \exp[\lambda f(H, M) + \overline{\lambda}\, \partial_M f(H, M)] \right) \tag{11}$$

where $H \equiv \sqrt{1-q}\, h - \sqrt{(1-q)x}\, t_0 - \sqrt{q}\, t$. In equation (11), we should take the minimum over λ and $\overline{\lambda}$ rather than the maximum because the integration over λ and $\overline{\lambda}$ runs along the imaginary axis in the complex plane, which should be deformed to pass the saddle point. When the saddle point happens to lie on the real axis, one may conveniently sweep along the real axis and the saddle point corresponds to the minimum point along the real axis.

Since $\overline{V}$ depends on s only through x, it is straightforward to show that $\overline{V}$ reaches its maximum at $x = 0$ and it follows that we can set $t_0 = 0$ in equation (11). Since q represents the typical correlations of the solution of equations (8), the typical fractional volume should shrink to zero as q approaches unity. Accordingly, the optimal storage capacity is then determined in this limit. When q approaches unity the last term in equation (11) diverges as $\sim (1-q)^{-1}$, and we write

$$\int \mathrm{D}t \log \int \mathrm{D}h \exp[\lambda f(H, M) + \overline{\lambda}\, \partial_M f(H, M)] \longrightarrow \frac{1}{1-q} \int \mathrm{D}t\, \Omega_t(H_t)$$

where the function $\Omega_t(H)$ is then given by

$$\Omega_t(H) \equiv -\tfrac{1}{2}(H+t)^2 + \lambda(1-q)f(H,M) + \overline{\lambda}(1-q)\,\partial_M f(H,M)$$

and H_t is the value of H leading to the maximum of Ω_t. Therefore g_2 also exhibits $(1-q)^{-1}$ divergence:

$$g_2 = \frac{1}{1-q}\max_{m\leqslant M\leqslant 1}\min_{\lambda,\overline{\lambda}}\left(-\lambda(1-q)M + |\overline{\lambda}(1-q)| + \int \mathrm{D}t\,\Omega_t(H_t)\right)$$

and the saddle-point equation over λ reads

$$M = \int \mathrm{D}t\, f(H_t, M) \tag{12}$$

where the dependence on the variables λ and $\overline{\lambda}$ is implicit through $H_t(\lambda, \overline{\lambda}, M; T)$. For the minimization over $\overline{\lambda}$, one should consider two cases. The first case is that the minimum occurs at $\overline{\lambda} = 0$. This happens when the absolute value of $\int \mathrm{D}t\,\partial_M f(H_t, M)$ with $\overline{\lambda} = 0$ and $\lambda = \lambda_0$, where λ_0 is given by the solution of equation (12) with $\overline{\lambda} = 0$, is less than unity. In the other case, the minimum occurs at $\overline{\lambda} \neq 0$ and the saddle point in the $(\lambda, \overline{\lambda})$ plane is given by the solution of equation (12) together with the equation

$$-\operatorname{sign}(\overline{\lambda}) = \int \mathrm{D}t\,\partial_M f(H_t, M).$$

In both cases, we denote the saddle point to be $(\lambda_0, \overline{\lambda}_0)$, and write g_2 in the form

$$g_2 = -\frac{1}{2(1-q)}\min_{m\leqslant M\leqslant 1}\alpha_0^{-1}(M;T)$$

where

$$\alpha_0^{-1}(M;T) \equiv \int Dt[t + H_t(\lambda_0, \overline{\lambda}_0, M; T)]^2. \tag{13}$$

Combining the above, we finally obtain the typical fractional volume:

$$\overline{V} = \exp\left(\frac{CN}{2(1-q)}\left[1 - \alpha\min_{m\leqslant M\leqslant 1}\alpha_0^{-1}(M;T)\right]\right)$$

which vanishes for

$$\alpha > \alpha_c \equiv \max_{m\leqslant M\leqslant 1}\alpha_0(M;T). \tag{14}$$

Interestingly, $\alpha_0(M;T)$ also represents the maximum storage capacity for the stationary value of the overlap M_μ^* in the range $M \leqslant M_\mu^* \leqslant M + \delta M$. Due to the mean-field nature of the network, the optimal storage capacity is given by the maximum value of α_0 for the given range of M.

Since $\alpha_0(M;T)$ defined in equation (13) involves minimization with respect to two variables $(\lambda, \overline{\lambda})$ in addition to the two Gaussian integrals over t and z (representing the thermal average), we computed them numerically. Figure 1 shows the typical behaviour of

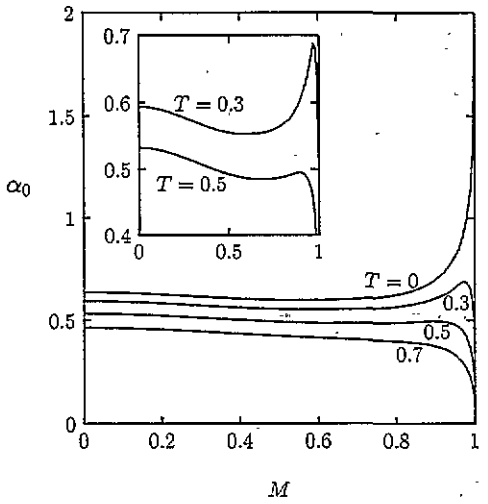

Figure 1. Typical behaviour of the maximum storage capacity α_0 as a function of the stationary overlap M^* at various temperatures: from top to bottom, $T = 0, 0.3, 0.5$ and 0.7, respectively. Detailed behaviour at $T = 0.3$ and 0.5 is displayed in the inset.

$\alpha_0(M; T)$ for several values of T, with the detailed behaviour for $T = 0.3$ and 0.7 displayed in the inset. There exist three types of M-dependence on $\alpha_0(M; T)$ according to T. When T is higher than $T_1 (\approx 0.566)$, the maximum capacity α_0 decreases monotonically with M. For $T < T_1$, α_0 exhibits a local minimum as well as a local maximum (as shown in the inset of figure 1). This local maximum (at non-zero M) is, in fact, the global maximum of α_0 for T lower than $T_2 (\approx 0.414)$ whereas α_0 reaches its maximum at $M = 0$ for $T_2 < T < T_1$.

In contrast to the naive expectation, the maximum storage capacity α_0 is not monotonic with M (or with the error allowed) when the temperature is lower than T_1. It is of interest to note that there are two types of fluctuation in the dynamics: one is the thermal fluctuations associated with the synaptic noise and controlled by the temperature T while the other is *the dynamical fluctuations* described by $\sqrt{1 - M^2}$. The latter fluctuations come from the distribution of states with definite overlap M and may be considered to be driven by the dynamics itself. In general, thermal fluctuations randomize spin orientations and tend to decrease the capacity whereas dynamical fluctuations affect the capacity in a more or less subtle manner because the level of dynamical fluctuations depends on the overlap. At zero temperature ($T = 0$) and for $M = 1$ neither thermal nor dynamical fluctuations are present. In this limit, perfect matching is allowed, leading to $\alpha_0 = 2$ similar to Gardner's result [3]. However, a small departure from $M = 1$ induces dynamical fluctuations in the potential of the neurons, so that the maximum storage capacity α_0 decreases rapidly as M is reduced. At $T = 0$, as shown in figure 1, α_0 reaches its maximum at $M = 1$ and hence we have the optimal storage capacity $\alpha_c = 2$ regardless of the tolerance parameter m. At finite temperatures thermal fluctuations always exist, which prohibits perfect matching. In this case it may be expected that allowing some error (i.e. $m < 1$) increases the capacity. On the other hand, reducing the overlap introduces dynamical fluctuations and eventually reduces the capacity if the temperature is not too high ($T < T_1$). Near $M \approx 0$, reduction in the overlap generally increases the capacity at any temperature since dynamical fluctuations favour small values of the overlap. Here we stress that one should not expect the divergence of the capacity in the limit $m \to 0$ because the trivial solution $M_\mu^* = 0$ is not included.

From the curves of α_0 it is straightforward to get the optimal storage capacity α_c defined

by equation (14) for given tolerance parameter and temperature. The typical behaviour of α_c as a function of m is shown in figure 2 at several temperatures. At temperatures higher than T_1 (≈ 0.566), α_0 is a monotonic decreasing function of m. Consequently, we have $\alpha_c(m; T) = \alpha_0(M = m; T)$, and the curves of α_c are identical to those of α_0. For $T < T_1$, there appears a plateau on which the optimal storage capacity is constant over some range of the tolerance parameter. (See figure 2. The boundary of this region is displayed by the dotted line.) For $T < T_2(\approx 0.414)$, it is interesting to note that α_0 reaches its maximum near $M \approx 1$ and that a large value of the overlap is mostly favoured.

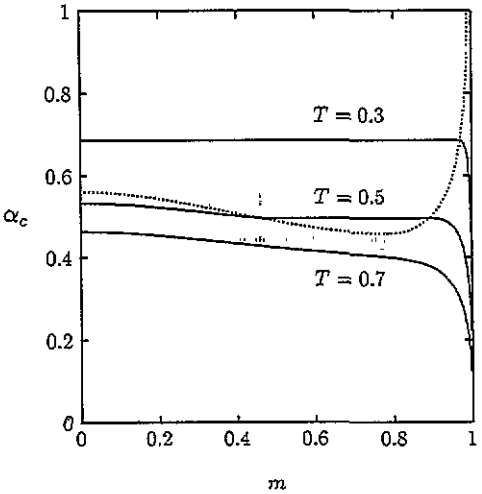

Figure 2. Typical behaviour of the optimal storage capacity α_c as a function of the tolerance parameter m at various temperatures: from top to bottom, $T = 0.3, 0.5$ and 0.7, respectively. The dotted curve shows the boundary of the region in which α_c is constant.

Consider a problem in which we want to store and recall αC random patterns in the network at temperature T at the best performance, that is we want the stationary overlap to be as large as possible. When α is small, one can easily find a set of couplings ($\{J_{ij}\}$) that yields the stationary value of the overlap near unity. As α increases, it becomes more difficult to find such a set of couplings. In general the quality of performance will deteriorate with the storage ratio α. Since $\alpha_0(M; T)$ is the maximum storage capacity with the stationary overlap M at temperature T, the best performance $M_p(\alpha; T)$ for given storage ratio α and temperature T is determined by the largest value of M for which $\alpha_0(M; T)$ is greater than α. This implies $\alpha = \alpha_c(M_p; T)$ and the curve of the best performance also corresponds to the optimal storage capacity. Therefore figure 2 also represents curves of the best performance with the abscissa and the ordinate denoting M_p and α, respectively. As the number of stored patterns increases, a first-order transition from good to poor performance occurs at temperatures not too high ($T < T_1$). Interestingly, at low temperatures ($T < T_2$) the network near saturation naturally favours high-quality performance; there are no networks yielding low-quality performance.

4. Analysis using approximate criterion

In this section, we study the proposed scheme with the approximate criterion given by equation (5) instead of equation (3) because it is very difficult to solve the dynamics of

neural networks in general. In fact even with this approximation the calculation is not easy and we implement the calculation only for extremely diluted neural networks. The validity of the approximation will be tested against the result of section 3. The fractional volume V of the space of interactions ($\{J_{ij}\}$) satisfying equations (4) and (5) is given by

$$V = \int \left[\prod_{i\neq j} \mathrm{d}J_{ij}\right] \prod_{\mu=1}^{\alpha C} \theta\left(\frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \tanh\left(\frac{\beta}{\sqrt{C}} \sum_j J_{ij}\xi_j^\mu\right) - m\right) \prod_i \delta\left(\sum_j (J_{ij})^2 - C\right)$$
$$\times \left\{ \int \left[\prod_{i\neq j} \mathrm{d}J_{ij}\right] \prod_i \delta\left(\sum_j (J_{ij})^2 - C\right)\right\}^{-1}.$$

Although there is no restriction on the correlations between J_{ij} and J_{ji}, different sites i and j are not decoupled because of equation (5); thereby it is not easy to calculate the typical fractional volume $\overline{V} \equiv \exp(\langle\langle \log V\rangle\rangle)$, which involves the average $\langle\langle\rangle\rangle$ of $\log V$ over the distribution of the random patterns $\{\xi^\mu\}$. Nevertheless the calculation can be performed for an extremely diluted network as discussed in the previous section. In this case the cumulant expansion can be cut-off as before at the second order.

Following a procedure similar to that in section 3, we obtain the typical fractional volume in the form

$$\overline{V} = \exp\left(CN\left[\frac{1}{2}\log(1-q) + \frac{1}{4}\log(1-x^2) + \frac{1}{2}\frac{q-rx}{(1-q)(1-x^2)} + \alpha g_2\right]\right) \tag{15}$$

where the function g_2 in this case is given by

$$g_2 \equiv \frac{1}{N}\int_{-\infty}^{\infty} \prod_{i=1}^{N} \mathcal{D}t_i \log\left[\int_{-\infty}^{\infty} \frac{\mathrm{d}t_0}{\sqrt{2\pi/N}} \exp(-\tfrac{1}{2}Nt_0^2)\right.$$
$$\left. \times \int \prod_{i=1}^{N} \mathrm{D}h_i\, \theta\left(\frac{1}{N}\sum_{i=1}^{N} \tanh[\beta(\sqrt{1-q}\,h_i - \sqrt{s-r}\,t_0 - \sqrt{q}\,t_i)] - m\right)\right]$$

with the same notation as in section 3. Using the integral representation of the θ-function

$$\theta\left(\frac{1}{N}\sum_{i=1}^{N}\tanh h_i - m\right) = \int_m^1 \mathrm{d}M \int_{-\mathrm{i}\infty}^{+\mathrm{i}\infty} \frac{\mathrm{d}\lambda}{2\pi\mathrm{i}/N} \exp\left(-N\lambda(M - \frac{1}{N}\sum_{i=1}^{N}\tanh h_i)\right)$$

and noting the range of the integral variable M, we obtain

$$g_2 = -m\lambda - \tfrac{1}{2}t_0^2 + \int \mathrm{D}t \log\left(\int \mathrm{D}h \exp\{\lambda \tanh[\beta(\sqrt{1-q}\,h - \sqrt{(1-q)x}\,t_0 - \sqrt{q}\,t)]\}\right)$$

where t_0 and λ are to be determined by the saddle-point equations. Since $\overline{V}$ depends on s only through x, it is straightforward to show that $\overline{V}$ reaches its maximum at $x = 0$ and $t_0 = 0$.

The optimal storage capacity can be determined according to the condition that the typical fractional volume shrinks to zero, which happens as q approaches unity. In this limit, the typical fractional volume given by equation (15) has the leading term:

$$\overline{V} = \exp\left(CN\left[\tfrac{1}{2}\log(1-q) - \alpha m\lambda + \frac{\alpha}{1-q}\int \mathrm{D}t\, \Omega_t(H_t)\right]\right)$$

where $\Omega_t(H) \equiv -\frac{1}{2}(H+t)^2 + \lambda(1-q)\tanh(\beta H)$ in this case and H_t is again the value of H leading to the maximum of Ω_t. Note that H_t depends on $\lambda(1-q)$ and T in addition to t. Thus, in the limit $q \to 1$ and $\lambda \to \infty$ with $\lambda(1-q)$ fixed, the saddle-point equations read as

$$\alpha_c = \left(\int \mathrm{D}t\{t + H_t[\lambda(1-q); T]\}^2 \right)^{-1} \tag{16a}$$

$$m = \int \mathrm{D}t \tanh(\beta H_t[\lambda(1-q); T]) \tag{16b}$$

where H_t is, by definition, given by the solution of the equation

$$\Omega_t'(H) \equiv -(H+t) + \lambda(1-q)\beta[1 - \tanh^2(\beta H)] = 0. \tag{17}$$

Equation (17) has a unique root unless $\lambda(1-q) > (3\sqrt{3}/4)T^2$ and $t_- < t < t_+$. (In this range equation (17) has three roots.) Here $t_\pm$ are defined to be

$$t_\pm \equiv \Omega_{t=0}'[H = -T\tanh^{-1}\{\tfrac{2}{\sqrt{3}}\cos(\tfrac{1}{3}(\pi \pm \phi))\}] \qquad \text{with} \qquad \phi \equiv \cos^{-1}\left(\frac{3\sqrt{3}T^2}{4\lambda(1-q)}\right).$$

At zero temperature it is straightforward to compute H_t and to write the optimal storage capacity in the form

$$\alpha_c = \left(\int_0^{\sqrt{2}\mathrm{erf}^{-1}(m)} \mathrm{D}t\, t^2 \right)^{-1}$$

while, at finite temperatures, equations (16) can be solved numerically.

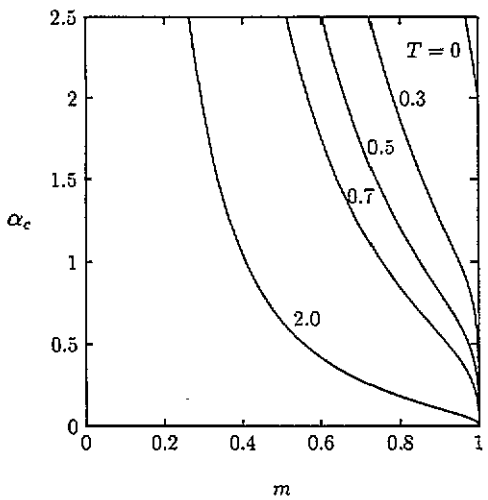

Figure 3. The optimal storage capacity α_c as a function of the tolerance parameter m at $T = 0, 0.3, 0.5, 0.7$ and 2.0 when the alternative approximate criterion given by equation (5) is used.

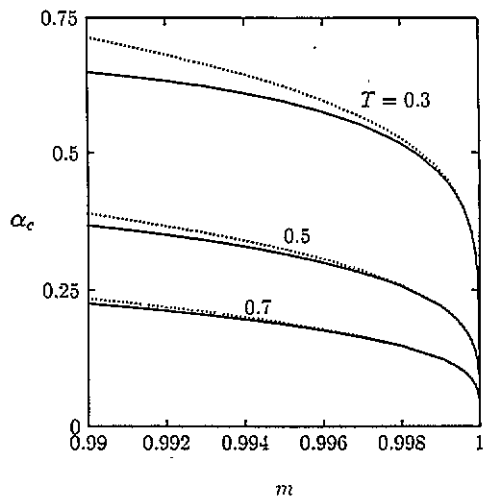

Figure 4. Detailed behaviour of the optimal storage capacity α_c for the tolerance parameter m near unity. Full and dotted curves are results of equations (3) and (5), respectively.

Figure 3 displays the optimal storage capacity α_c as a function of the tolerance parameter m at various temperatures. The overall dependence of α_c on m is qualitatively different from that obtained in section 3. In particular α_c diverges as $m \to 0$. Our approximate criterion loses its validity near $m = 0$ as it should. However, m-dependence of α_c near $m = 1$ is not too disparate from that of figure 2 at finite temperatures. The α_c curves at various temperatures shown in figure 3 are reproduced to expose the detailed behaviour near $m = 1$ in figure 4, which, for comparison, also displays the corresponding curves obtained in section 3. At given temperature the two curves indeed coincide with each other in the limit $m \to 1$. Therefore we conclude that the approximate scheme based on equation (5) is valid for m close to unity.

It is of interest to apply the above scheme to the dynamic model of neural networks [12], where a neuron is forced to have state $s_i = -1$ during the refractory period. As a consequence, Gardner's method cannot be applicable even at zero temperature. In the dynamic model, equations describing the time evolution of relevant physical quantities generally assume the form of differential-difference equations due to the retardation in interactions. In particular, the activity $\langle s_i(t)\rangle$ of the ith neuron at time t and the overlap $M_\mu(t)$ between the network state and the μth pattern at time t are determined by the differential-difference equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle s_i(t)\rangle = (\tfrac{1}{2}-a) - (\tfrac{1}{2}+a)\langle s_i(t)\rangle + \tfrac{1}{2}(1-\langle s_i(t)\rangle)\tanh\left(\frac{\beta}{\sqrt{N}}\sum_j J_{ij}\langle s_i(t-1)\rangle\right)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}M_\mu(t) = -(\tfrac{1}{2}+a)M_\mu(t) + \frac{1}{2N}\sum_{i=1}^{N}\xi_i^\mu(1-\langle s_i(t)\rangle)\tanh\left(\frac{\beta}{\sqrt{N}}\sum_j J_{ij}\langle s_i(t-1)\rangle\right)$$

respectively. Here a represents the ratio of the refractory period to the time duration of the action potential. In the stationary state, the overlap M_μ takes the form

$$M_\mu = \frac{4a}{1+2a}\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i^\mu\tanh((\beta/\sqrt{N})\sum_j J_{ij}\langle s_i\rangle)}{1+2a+\tanh((\beta/\sqrt{N})\sum_j J_{ij}\langle s_i\rangle)}. \tag{18}$$

Since $\xi_i^\mu\sum_j J_{ij}\xi_j^\mu$ tends to be positive for typical types of interactions, we may, in the extreme-dilution limit, make an approximation in equation (18) as

$$M_\mu = \frac{1}{N}\sum_{i=1}^{N}\frac{4a\tanh((\beta/\sqrt{N})\sum_j\xi_i^\mu J_{ij}\xi_j^\mu)}{(1+2a)^2-\tanh^2((\beta/\sqrt{N})\sum_j\xi_i^\mu J_{ij}\xi_j^\mu)}.$$

The optimal storage capacity of the dynamic model is now determined by equations (16) except for that the $\tanh x$ function is replaced by $4a\tanh x/[(1+2a)^2-\tanh^2 x]$. Unlike the Hopfield model which discretizes the time, the dynamic model takes into account the existence of relevant time scales and, consequently, displays the overlap $M_\mu = 1/(1+a)$ in the case of perfect recall. This is reflected in the equation corresponding to equation (16*b*). For comparison with the Hopfield model, therefore, we rescale the tolerance parameter m by $\tilde{m} \equiv (1+a)m$, and finally get

$$\alpha_c(\tilde{m}) = \left(\int_0^{\sqrt{2}\,\mathrm{erf}^{-1}(\tilde{m})}\mathrm{D}t\,t^2\right)^{-1}$$

at zero temperature.

5. Discussion

We have proposed a new method for studying the optimal storage property of neural networks at finite temperatures and investigated the optimal storage capacity α_c for an extremely diluted network as a function of temperature T and tolerance parameter m. At zero temperature, it has been found that $\alpha_c = 2$ regardless of the tolerance parameter whereas at finite temperature α_c vanishes in the perfect matching limit ($m \to 1$), in general increasing with the tolerance. The best performance for given storage ratio α and temperature has also been obtained. At low temperatures ($T < T_1 \approx 0.566$) the network exhibits a first-order transition from high- to low-quality performance as the number of stored random patterns is increased. High-quality performance seems to be naturally favoured by extremely diluted networks if the level of noise is not too high. We have also studied an approximate scheme, which yields qualitatively different results except near $m = 1$. The crude approximation $\langle s_i \rangle \approx \xi_i^\mu$ used in equation (5) has been found not to be so good in the whole range of m; for m close to unity, however, it can be regarded as an accurate approximation.

Instead of the proposed criterion for recall, one may consider a slightly different criterion: the time average of $\xi_i^\mu s_i(t)$ *for each site* i should be greater than m. In the same spirit as that in equation (5), one may consider the problem of calculating the typical fractional volume of the space of interactions satisfying equation (4) and

$$\xi_i^\mu \tanh\left(\frac{\beta}{\sqrt{N}} \sum_j J_{ij} \xi_j^\mu\right) > m \tag{19}$$

for each i. The problem is then equivalent to Gardner's problem with her parameter κ given by $\kappa = T \tanh^{-1} m$, which implies that the required basin of attraction grows larger with the level of synaptic noise and with the accuracy of recalling. At zero temperature this leads to the optimal capacity $\alpha_c = 2$ regardless of m, whereas at finite temperatures α_c increases with the tolerance. Despite their resemblance, the behaviour of the optimal storage capacity with the criterion (19) is qualitatively different from that of (5). Although the argument of tanh in both cases is a sum over many sites j, it should be strongly correlated with ξ_i^μ if the network ($\{J_{ij}\}$) is to function as associative memory. In general, the overlap on site i, given by the right-hand side of equation (19), will vary from site to site according to some distribution with finite variance. Criterion (5) in the large N limit requires the overlap averaged over that distribution to be greater than m while that of (19) demands that the overlap be greater than m for each site.

Finally, there are several points for further investigation. Since we have assumed the replica- and site-symmetric ansatz in the calculation of the typical fractional volume, its stability against replica-symmetry-breaking should be checked. It would also be of interest to extend our results to the fully connected networks and other types of networks.

Acknowledgments

This work was supported in part by the Korea Science and Engineering Foundation through the Centre for Theoretical Physics, Seoul National University.

References

[1] For a recent review and an extensive list of references, see, e.g., Müller B and Reinhardt J 1990 *Neural Networks: An Introduction* (Berlin: Springer) and the articles in 1989 *J. Phys. A: Math. Gen.* **22**

[2] Mézard M, Parisi G and Virasoro M A 1986 *Spin Glass Theory and Beyond* (Singapore: World Scientific)

[3] Gardner E 1987 *Europhys. Lett.* 4 481; 1988 *J. Phys. A: Math. Gen.* **21** 257

[4] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271

[5] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983

[6] Krauth W and Opper M 1989 *J. Phys. A: Math. Gen.* **22** L519

[7] Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **23** 2613

[8] Gardner E, Gutfreund H and Yekutieli I 1989 *J. Phys. A: Math. Gen.* **22** 1995

[9] Engel A 1990 *J. Phys. A: Math. Gen.* **23** L285

[10] Bauer K and Krey U 1991 *Z. Phys.* B **84** 131

[11] Kanter I 1992 *Europhys. Lett.* **17** 181

[12] Choi M Y 1988 *Phys. Rev. Lett.* **61** 2809
Shim G M, Choi M Y and Kim D 1991 *Phys. Rev.* A **43** 1079

[13] Mézard M 1989 *J. Phys. A: Math. Gen.* **22** 2189

[14] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* 4 167

[15] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969

[16] Amit D J, Evans M R, Horn H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361

[17] Keppler T B and Abbott L F 1988 *J. Physique* **49** 1657