Hidden information in Hamiltonian systems with discrete weights

Ido Kanter
Department of Physics, Bar-Ilan University, Ramat Gan 52900, Israel
(Received 14 March 1991)

Statistical mechanics is applied to estimate the maximal capacity per weight (α_c) of a network consisting of N binary units and NC binary weights, where C is the average connectivity. For some range of η, which measures the average correlation between J_{ij} and J_{ji}, the replica-symmetric solution gives the exact α_c, where in the symmetric limit the capacity per bit is greater than 1. However, for symmetric systems with deeper weight depths, the capacity per bit is less than 1, and no perfect information engine exists.

PACS number(s): 87.10.+e

Much work has been done in the recent past to estimate the computational capability of various neural network architectures. The basic task of these networks is to store random patterns, configurations that are fixed points of the dynamics. The notion that played an important role in estimating the computational ability of the network was the capacity per weight (synapse), the ratio between the total number of bits of the patterns, and the total number of weights [1,2]. It was recently found that it is constructive to consider, in addition to the capacity per weight, the capacity per bit. The definition of this notion is the ratio between the number of bits of the patterns and the necessary number of bits for the representation of the weights. It will be proved in this work that, surprisingly, neither of these notions is the exact notion for estimating the computational ability of the network, since even the capacity per bit can be greater than 1. The discussed results are derived by the powerful method suggested by Gardner [1], which was mostly applied to asymmetric networks. Nevertheless, the computational capability of symmetric networks, which are the physical ones, is still in question and is at the center of our discussion.

The discussed network consists of N binary units $s_i=\pm 1$, coupled through binary weights $J_{ij}=\pm 1$. The weights are correlated [3] by the symmetry order parameter η

$$\eta=\sum_{\substack{i=1\\ i\neq j}}^{N} J_{ij}J_{ji}/C \ , \tag{1}$$

where C is the average connectivity. Each unit is connected on the average to C random units with binary weights. Therefore, the existence of a weight J_{ij} implies also the existence of the weight J_{ji}. In the limits $\eta=-1$ and 1 the matrix is antisymmetric and symmetric, respectively, whereas for $\eta=0$ there are no correlations between J_{ij} and J_{ji}. The analytical part of this work concentrates only on the highly diluted limit, where $C=0(\ln(N))$.

The dynamical evolution of the network is, according to zero-temperature Monte Carlo dynamics,

$$s_i(t+1)=\mathrm{sgn}\left[\sum_{j\ (\neq i)}\frac{J_{ij}s_j(t)}{\sqrt{C}}-\kappa\right] , \tag{2}$$

where κ is the stability parameter [1,4]. The task of the network is to store P random patterns $\xi_i^\mu=\pm 1$ with equal probability, where $i=1,\ldots,N$, $\mu=1,\ldots,P$, and α is defined to be equal to P/C. The problem is to calculate the maximal number of patterns P_c that can be taught to the network as a function of the symmetry parameter η.

This model with continuous weights has been addressed in Ref. [3] and solved within the replica-symmetric (RS) ansatz. The capacity was studied as a function of η and κ, and in particular it was found that the maximal storage capacity per weight, $P/c=2$, is obtained at $\eta=1/\pi$.

Before going through the details, let us first summarize the main results of this work. (a) The maximal capacity at $\eta=1$ is equal to ~ 0.56. This result indicates that in contrast to asymmetric systems, the capacity per bit in symmetric systems could be grater than 1. (b) The maximal storage capacity has a maximum at $\eta_{\max}\sim 0.33$, where $\alpha_c\sim 0.83$, which is identical to the binary perceptron [5]. (c) At the maximal capacity and for $\eta<\eta_{\max}$ the system is characterized by a finite entropy, but the average correlation h between J_{ij} and J_{ji} is equal to η. At $\eta_{\max}$, the entropy vanishes and $h=\eta_{\max}$. However, for $\eta>\eta_{\max}$ the entropy vanishes but $h<\eta$. Hence, the maximal capacity is characterized by one of the following two criteria: zero entropy or (and) $h=\eta$. (d) At finite temperature and $\eta>\eta_{\max}$ there is a critical line $\alpha_c(T_c,\eta,\kappa)$ where the entropy of the system vanishes. For $T<T_c$, the functional order parameter of the glassy phase is characterized by $q_0=q_{\mathrm{RS}}, q_1=1$ and the breakpoint $m=T/T_c$.

The derivation of the results is based on the method suggested by Gardner and Derrida [4]. For each set of coupling the energy per unit is defined as the number of patterns that are not embedded in this unit

$$E_i(J)=\sum_{\mu=1}^{p}\Theta\left[-\xi_i^\mu\sum_{j\ (\neq i)}\frac{J_{ij}\xi_j^\mu}{\sqrt{C}}\right] . \tag{3}$$

Following Ref. [4], for a given set of patterns we introduce the partition function at temperature $T=1/\beta$,

$$Z=\mathrm{Tr}_{J_{ij}=\pm 1}\exp\left[-\beta\sum_i E_i\right] . \tag{4}$$

In the zero-temperature limit the partition function, Eq. (4), is exactly the volume V (the number of configurations), which achieves the desired output, and the entropy per weight is $\ln(Z)/NC$. The patterns can be learned if and only if the zero-temperature internal energy vanishes. It is obvious that the entropy is a nonincreasing function of α, but it is unnecessary that the maximal capacity be defined as the lowest α where the entropy vanishes.

A nontrivial upper bound for the maximal capacity is obtained by the annealed approximation, where the free energy is given by $-T\ln\langle Z\rangle/NC$, where Z is defined in Eq. (4) and $\langle\ \rangle$ stands for the average over the disorder. In the zero-temperature limit one can find that the entropy per weight is given by

$$\ln V/NC=\tfrac{1}{2}\{D+\ln[4\cosh(D)]\} -\alpha[t_0^2/2+\ln H(\sqrt{\eta}t_0)] , \tag{5}$$

where $H(x)=\int_x^\infty Dx$, $Dt=e^{-t^2/2}dt/(2\pi)^{1/2}$, and t_0 and D are fixed by their saddle-point (SP) equations, $\partial V/\partial D=\partial V/\partial t_0=0$. The result for $\alpha_{\mathrm{ann}}(\eta)$ at the critical capacity is given in the upper curve of Fig. 1. The result that $\alpha_{\mathrm{ann}}(0)=1$ is not surprising, since in this limit each weight is an independent variable that contains 1 bit of information. Nevertheless, the result that $\alpha_{\mathrm{ann}}>1$ for $0<\eta<0.75$ and that α_{ann} has a maximum at $\eta\sim 0.42$ is surprising. Furthermore, as $\eta\to 1$, the number of independent weights is $NC/2$, but the annealed approximation gives a critical capacity ~ 0.7, which is greater than 0.5. Nevertheless, the exact maximal capacity and the nature of the system are investigated using the replica method [6].

In the computation of extensive thermodynamic quantities one concentrates on $\ln Z$, which is averaged over the distribution of the quenched random patterns $\{\xi_i^\mu\}$, using the replica method. In this calculation one introduces two types of order parameters,

$$q^{\alpha\beta}=\frac{1}{C}\sum_{j\ (\neq i)} J_{ij}^\alpha J_{ij}^\beta , \tag{6}$$

$$h^{\alpha\beta}=\frac{1}{C}\sum_{j\ (\neq i)} J_{ij}^\alpha J_{ji}^\beta , \tag{7}$$

which are assumed to be independent of the unit i. The physical meaning of $q^{\alpha\beta}$ is the overlap between the weights in two different replicas, α and β, and $h^{\alpha\beta}$ measures the correlations between J_{ij} and J_{ji} in two different replicas. Under the RS assumption, $q^{\alpha\beta}=(1-q)\delta_{\alpha\beta}+q$ and $h_{\alpha\beta}=(\eta-h)\delta_{\alpha\beta}+h$ one can show that

$$G_{\mathrm{RS}}=\langle\langle \ln Z\rangle\rangle=ext_{\{q,h,t_0,F,H,D\}}\left[-\tfrac{1}{2}[F(1-q)+Hh+\eta D]+\tfrac{1}{2}\langle\ln[2e^{-D-H}\cosh(X_1)+2e^{D+H}\cosh(X_2)]\rangle_t-\frac{\alpha t_0^2}{2} +\alpha\int_{-\infty}^{\infty}Dt\ln[H(X_3)]\right] , \tag{8}$$

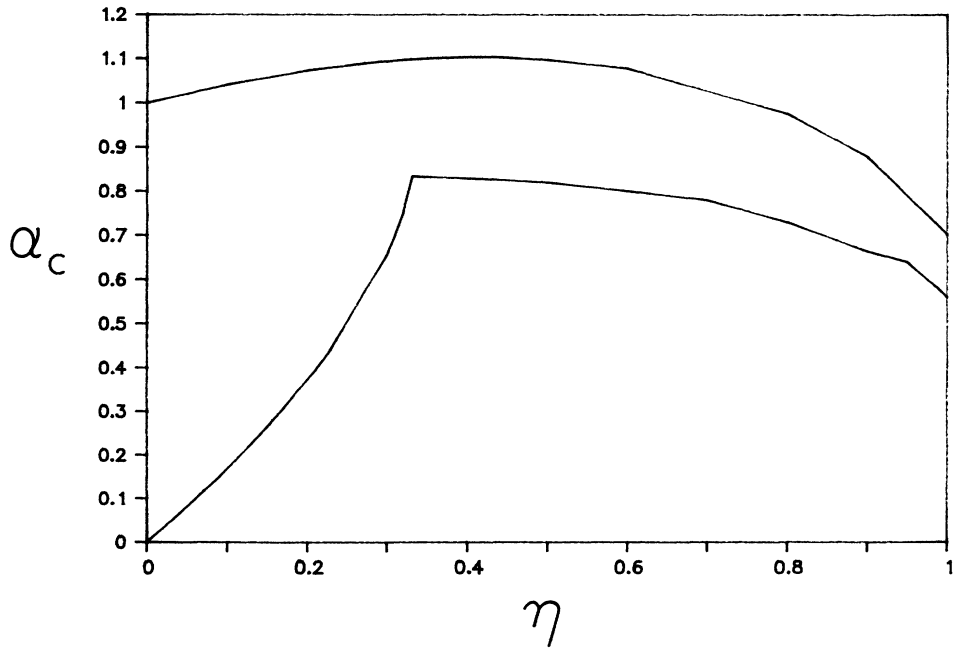

FIG. 1. α_c as a function of η. The upper curves is the annealed approximation and the lower curve is the RS solution.

where the symbol $\langle\langle\ \rangle\rangle$ stands for the average over the patterns,

$$X_1=\sqrt{F-H}\,(t_1+t_2)+2\sqrt{H}\,t_3\ ,$$

$$X_2=\sqrt{F-H}\,(t_1+t_2)\ ,$$

$$X_3=(\kappa+\sqrt{\eta-h}\,t_0+\sqrt{q}\,t)/\sqrt{1-q}\ ,$$

$\langle\ \rangle_t$ stands for the average over the Gaussian variables t_1, t_2, and t_3, and the six parameters are fixed by the SP equations. The explicit form of the SP equations will be given elsewhere [7]; however, let us only mention a technical point that exists in the following cases: for the general case, the limit $h=\eta$ and the limit $\eta=1$, needs to only solve 4, 3, and 2 coupled equations, respectively.

The solution of the SP equations at zero temperature and $\kappa=0$ indicates that one should distinguish between the following two regimes. (a) $0\le\eta\le\eta_{\max}\sim0.33$, the maximal storage capacity increases from $\alpha_c=0$ at $\eta=0$ up to ~0.83 at $\eta_{\max}$ (see Fig. 1). For a fixed η, the order parameters q and h increase as a function of α, where at the criticality $h=\eta$ but $q<1$. The entropy decreases with α, but at the criticality G is positive and decreases to zero at $\eta_{\max}$. (b) $\eta_{\max}<\eta\le1$, the maximal capacity decreases from ~0.83 at $\eta_{\max}$ up to ~0.56 at $\eta=1$ (see Fig. 1). At the criticality, the entropy G vanishes, but $h<\eta$ and $q<1$.

It is important to note that the maximal storage capacity is obtained at $\eta_{\max}$, where J_{ij} and J_{ji} are correlated and $\alpha_c(\eta_{\max})$ is equal to the maximal capacity of the binary perceptron [5]. There is a similarity between the solution of the annealed approximation and the RS solution. In both cases $\alpha_c(\eta)$ has a maximum at some intermediate value of η; however, the RS solution decreases this value from 0.42 to 0.33. The result $\eta_{\max}=0.33$ for the binary case indicates that the maximal capacity that is obtained at the symmetry order parameter $\eta_{\max}$ is not sensitive to the type of the weights, since for continuous weights $\eta_{\max}=1/\pi\sim0.32$ [2]. However, in the binary case there are no solutions with $\eta<0$.

The RS solution indicates two surprising results. The maximal storage capacity per bit at $\eta=1$ is 1.12, since $\alpha_c>0.5$ and the number of independent weights is only $NC/2$. The second result is that, at $\alpha_c(\eta)$, G can be greater than zero. In previous analytical solutions, such as the perceptron with continuous weights [1,4], the maximal capacity is obtained where $G\to0$ and $q\to1$. The critical capacity of the binary perceptron [5] is obtained at $G\to0$ but with $q<1$. Our results indicate that at least within the RS solution the third possibility also exists, and the critical storage capacity is obtained where $G>0$, $q<1$, and $h=\eta$. In order to check whether the RS solution gives the exact result, we examine our network in the replica-symmetry-breaking (RSB) scheme [6].

The average $\ln Z$ with one step of RSB is given by

$$\begin{aligned}G_{1\,\mathrm{step}}&(q_0,q_1,F_0,F_1,h_0,h_1,H_0,H_1,D,t_0,V_0,m)\\&=\tfrac{1}{2}\{F_1[q_1(1-m)-1]+mF_0q_0+H_1h_1(1-m)+mH_0h_0+\eta D\}+\frac{1}{2m}\langle\ln\langle X^m\rangle\rangle_t\\&\quad+\alpha\left[-\frac{t_0^2}{2}-\frac{v_0^2}{2m}+\frac{1}{m}\int_{-\infty}^{\infty}Dy\ln\left[\int_{-\infty}^{\infty}Dz\,H^m(Y)\right]\right],\end{aligned}\tag{9}$$

where

$$\begin{aligned}X=&2e^{-(D+H_1)}\cosh[\sqrt{F_0-H_0}(t_1+t_2)+2\sqrt{H}y_1+\sqrt{\Delta F-\Delta H}\,(y_1+y_2]\\&+2e^{D+H_1}\cosh[\sqrt{F_0-H_0}(t_1-t_2)+\sqrt{\Delta F-\Delta H}\,(y_1-y_2)]\ ,\end{aligned}\tag{10}$$

$Y=(\sqrt{\eta-h}\,t_0+\sqrt{\Delta q}\,z+\sqrt{q_0}y+\sqrt{\Delta h}\,v_0)/\sqrt{1-q_1}$, $\Delta q=q_1-q_0$, $\Delta h=h_1-h_0$, $\Delta F=F_1-F_0$, $\Delta H=H_1-H_0$, and the symbols $\langle\ \rangle_y$ and $\langle\ \rangle_t$ stand for the average over the Gaussian variables y_i and t_i, $i=1,2,3$, respectively. In this parametrization the order parameters q_0 and h_0 represent the overalp and the correlation between two different valleys,

$$q_0=\frac{1}{C}\left\langle\!\!\left\langle\sum_j\langle J_{ij}\rangle_\alpha\langle J_{ij}\rangle_\beta\right\rangle\!\!\right\rangle\ ,\tag{11}$$

$$h_0=\frac{1}{C}\left\langle\!\!\left\langle\sum_j\langle J_{ij}\rangle_\alpha\langle J_{ji}\rangle_\beta\right\rangle\!\!\right\rangle\ ,\tag{12}$$

and q_1 and h_1 represent the self-overlap and the self-correlation within one valley,

$$q_1=\frac{1}{C}\left\langle\!\!\left\langle\sum_j\langle J_{ij}\rangle_\alpha^2\right\rangle\!\!\right\rangle\ ,\tag{13}$$

$$h_1=\frac{1}{C}\left\langle\!\!\left\langle\sum_j\langle J_{ij}\rangle_\alpha\langle J_{ji}\rangle_\alpha\right\rangle\!\!\right\rangle\ .\tag{14}$$

The breakpoint m relates to the Gibbs weight P_α of each one of the valleys, $1-m=\langle\langle\sum_\alpha P_\alpha^2\rangle\rangle$. The 12 parameters of Eq. (9) are fixed by their SP equations.

The investigation of these equations is a very heavy numerical task, since there are many parameters and the SP equations are constructed from six integrals. Nevertheless, in the limit $\eta\to1$, one can show that $D\to-\infty$, $h_1=q_1$, $h_0=q_0$, and $H_0=H_1$. Therefore, in this limit the

problem is reduced to the solution of only four coupled equations where each one of them is constructed from only five integrals. The SP equations in this limit were examined numerically and we find no solution except the RS solution. At finite temperature, the limit $q_1 \to 1$ is studied [5]. In this limit $F_1 \to \infty$ and one can verify that

$$G_{1\,\mathrm{step}}(q_0,F_0,1,\infty,H_1,t_0,m,\beta) = \frac{1}{m} G_{\mathrm{RS}}(q_0,m^2F,m_2H,t_0,\beta m)\ , \quad (15)$$

where G_{RS} is given by Eq. (8). The SP equations with respect to q_0, F, H, and t_0 imply that $t_0 = t_{0,\mathrm{RS}}$, $q_0 = q_{\mathrm{RS}}$, $F_0 = m^2 F_{\mathrm{RS}}$, and $H_1 = m^2 H_{\mathrm{RS}}$, where q_{RS}, $t_{0,\mathrm{RS}}$, F_{RS}, and H_{RS} are the RS order parameters at the inverse temperature βm. The SP with respect to m implies $S_{\mathrm{RS}} = S_{1\,\mathrm{step}} = 0$ where $\beta m = 1/T_c$, which is the inverse temperature where RS entropy vanishes. Therefore, in the zero-temperature limit and $\eta = 1$, the RS solution gives the *exact* solution. At finite temperature and $T < T_c$ there exists a one-step solution characterized by $q_0 = q_{\mathrm{RS}}$, $h = h_{\mathrm{RS}}$, and $q_1 = 1$, and the free energy is independent of T and is equal to the RS free energy at T_c. The one-step solution is first order in the sense that the order-parameter function $q(x)$ is discontinuous, while the free energy is still continuous. Such behavior was previously found in a few models such as the random-energy model, simplest spin glass [8,9] model, and the parity machine with binary weights [10] model.

The behavior of the network in the cases $\eta = 1$ and $\eta_{\max}$, which is identical to the binary perceptron, is similar. In both cases the RS solution gives the exact maximal storage capacity, where at finite temperature a one-step RSB solution appears below the temperature where the RS entropy vanishes. This result strongly indicates that the same behavior occurs for any η in the intermediate interval, $\eta_{\max} < \eta < 1$.

The investigation of the one-step RSB in the interval $0 < \eta < \eta_{\max}$ is a heavy numerical task that certainly warrants further study. Nevertheless, we cannot rule out the possibility that the maximal storage capacity is achieved at finite entropy where $G > 0$.

The result that the capacity per bit for $\eta = 1$ is 1.12 indicates a crucial difference between symmetric and asymmetric systems. In all previously examined feedforward asymmetric systems, and apparently for all of them, the maximal embedding information per bit is ≤ 1, but in symmetric systems it would be greater than 1. We indicate this additional information as hidden information. The general belief that the notions of information and capacity are identical is incorrect. These two notions are the same in the case of feedforward networks, but differ in the case of symmetric interactions where two weights are counted as one. In order to convince ourselves that this is indeed true, let us now give a simple example. The system consists of two spins that are connected by a binary symmetric weight. It is obvious that the maximal capacity is one pattern (two bits) and therefore the capacity per bit is 2. Nevertheless, a more physical system is a network consisting of a deeper weight depth $(J_{ij} = \pm 1, \ldots, \pm l)$. For the continuous case, α_c was found to be ~ 1.28 [3]. Therefore, for depth $l \geq 3$, and apparently even for $l = 2$, the maximal embedding information per bit is less than 1. This is another aspect of the hidden information. The stable points of a physical system with nontrivial interactions of depth l do not contain all the embedded information in the system. Therefore, an observer or a user of a physical process, at zero temperature, cannot use all the bits of information that are embedded in the system. One can consider such a system as an "engine," whose action is to convert (or to transfer) one type of information to another. In our case the two types are the presentation of the information in the space of the weights and the presentation of the information in the space of the units. Our results suggest a general rule that the efficiency of such a process is less than 1 for any nontrivial system. It is *impossible* to construct a *perfect information engine.* Note, that our process occurs at a fixed temperature and hence this rule cannot be derived directly from the basic thermodynamic laws.

The results of this work and their applications even to nondiluted symmetric systems were examined in simulations of fully connected symmetric networks. Since no learning algorithm is known, a stochastic algorithm is used. For each pattern, the stability of each one of the units is checked sequentially. A weight that is responsible for fixing a unit antiparallel to the pattern is flipped with probability p. Our simulations suggest that $p \sim 0.9$ gives the minimal covergence time. It is obvious that the convergence time grows exponentially with the number of weights. Nevertheless, for small systems one can verify self-consistently that the maximal number of steps allowed in our simulations is greater than ten times the average convergence time. Therefore, a possible effect of fluctuations in the convergence on α_c is small. For each P and N the fractional number of samples with a solution is obtained by averaging over at least 50 samples and α_c is fixed as the capacity where a solution is found with probability $\frac{1}{2}$. Our results for $N = 2$, 4, 6, 8, 10, 12 indicate that the maximal capacity is ~ 2, 1.33, 1.21, 1.18, 1.17, respectively, where $\alpha \equiv 2P/(N-1)$ and our theoretical prediction is $\alpha_c = 1.12$. These results suggest that the maximal storage capacity of a fully connected network is the same as for diluted networks and is very close to our theoretical prediction even for small systems.

Finally, note that the maximal capacity per bit is a continuous function of η as the capacity itself. One can show [7] that this quantity is given by $2\alpha_c / \{2 - [(1+\eta)\ln(1+\eta) + (1-\eta)\ln(1-\eta)]/(2\ln 2)\}$, which stands for the ratio between the number of necessary bits for the representation of the patterns and the number of necessary bits for the representation of the total number of configurations in the space of the weights. This quantity increases from zero at $\eta = 0$ up to 1.12 at $\eta = 1$, where it is greater than 1 for $\eta > 0.75$. However, unlike the case $\eta = 1$, for $\eta < 1$ it is not obvious how to encode the legal configurations in a simple way such that the capacity per bit is greater than 1.

I would like to thank E. Eisenstein for critical comments on the manuscript. This research is supported by the Israel Ministry of Science and Development.

[1] E. Gardner, J. Phys. A **21**, 257 (1988).
[2] See, for instance, the memorial volume of Elizabeth Gardner, J. Phys. A **22** (1989).
[3] E. Gardner, H. Gutfreund and I. Yekutieli, J. Phys. A **22**, 1995 (1989).
[4] E. Gardner and D. Derrida, J. Phys. A **21**, 271 (1988).
[5] W. Krauth and M. Mezard, J. Phys. (Paris) **50**, 3057 (1989).
[6] For a review, see M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[7] I. Kanter (unpublished).
[8] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[9] D. J. Gross and M. Mezard, Nucl. Phys. B **240**, 431 (1984).
[10] E. Barkai and I. Kanter, Europhys. Lett. **14**, 107 (1991).