Storage capacity and generalization error for the reversed-wedge Ising perceptron

G.J. Bex, R. Serneels, and C. Van den Broeck
Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

(Received 22 March 1994)

Using the replica formalism, we evaluate the storage capacity and the generalization error of a perceptron with a reversed-wedge transfer function and binary synaptic weights. Remarkably, both the storage capacity and the generalization threshold saturate the information theoretic (respectively upper and lower) bound $\alpha = 1$ for a specific choice of the width of the reversed wedge, suggesting that this perceptron may be an interesting building block for neural networks.

PACS number(s): 87.10.+e, 05.50.+q

During the last decade, the interest in neural networks has grown considerably. The questions at hand—memory, information processing, learning, and cognition—are of great fundamental importance, being related to the very nature of intelligence, but possess at the same time an enormous technological potential. This is so because neural networks hopefully share some of the more desirable properties of the human brain, such as a high degree of parallel operation, robustness, and flexibility. From the theoretical point of view, the attention has been focused on simpler networks, such as Hopfield networks or perceptrons which serve as building blocks for more complicated networks, and for which an analytical treatment becomes possible [1–3]. A particularly well-studied example is the case of the Ising perceptron [4–11], which is a normal perceptron but with the restriction that its weight vector components take only the values ±1. Both storage capacity and generalization error have been studied in the limit where the number of training examples p and the dimension of the input space N go to infinity with a fixed value of their ratio $\alpha = p/N$. Krauth and Mézard [6] obtained $\alpha \approx 0.83$ for the storage capacity from a one-step replica symmetry breaking calculation. This result was confirmed by simulations carried out by Krauth and Opper [7]. Györgyi [8] (see also [5,9]) studied the generalization error of an Ising perceptron learning examples generated by an Ising teacher perceptron and showed that a discontinuous transition to perfect learning (zero generalization error) takes place at $\alpha \approx 1.245$. These α values have to be compared with the value $\alpha = 1$, which is at the same time an upper bound for the storage capacity (since one component of the Ising weight vector can store at most one bit of information) and a lower bound for the threshold to zero generalization error (since one training pattern can convey at most one bit of information about the teacher). The fact that these bounds are not realized may at first look like the price to be paid, on the one hand, for the parallel and distributed character of the perceptron as a memory device and, on the other hand, for the fact that the information carried by randomly chosen examples will always exhibit a certain amount of redundancy. In the present paper, we show that this need not be so, and report on a simple variant of the perceptron for which both information theoretic bounds are actually saturated. This surprising result shows that the maximum rate of one bit information transfer per example up to full capacity of the network can be achieved in a distributed memory as well as in a process of learning from random examples.

The model under consideration is the Ising version of the so-called reversed wedge perceptron [12–14]. It returns the following binary output classification ξ_o on an N-dimensional input pattern $\vec{\xi}$:

$$\xi_o = \operatorname{sgn}\left[g\left(\frac{\vec{J}\cdot\vec{\xi}}{\sqrt{N}}\right)\right] \tag{1}$$

with

$$g(x) = (x+K)x(x-K)\ . \tag{2}$$

The weight vector $\vec{J}$ is taken to be of the Ising type $J_i = \pm 1, i = 1,\ldots,N$. Also, all the patterns will be assumed to be normalized as $|\vec{\xi}|^2 = N$. For the value $K = 0$, one recovers the familiar signum transfer function of the normal perceptron. A choice $K > 0$ corresponds to the insertion of a wedge of width $2K$ around $x = 0$ or, geometrically speaking, to the introduction of two extra hyperplanes, parallel and on both sides of the hyperplane orthogonal to $\vec{J}$ through the origin. Hence the reversed wedge perceptron can perform classifications, such as the XOR, which need not be linearly separable.

Let us first address the storage capacity problem of the Ising reversed wedge perceptron. The idea is to find an Ising vector $\vec{J}$ such that it correctly reproduces the classification ξ_o^μ for a set $\vec{\xi}^\mu$ of patterns $\mu = 1,\ldots,p$. These patterns and their corresponding classifications are assumed to be chosen at random and independent of each other. The number of Ising vectors that are compatible with these classifications is clearly given by the following expression:

$$\Omega = \sum_{\vec{J}=\{\pm1\}^N}\prod_{\mu=1}^{p}\theta\left(g\left(\frac{\vec{J}\cdot\vec{\xi}^\mu}{\sqrt{N}}\right)\xi_o^\mu\right) \tag{3}$$

with $\theta(x)$ the Heaviside function. Ω is of course a random number. However, in the limit $N \to \infty$ and $p \to \infty$ with the ratio $\alpha = p/N$ fixed, one expects that $\ln\Omega$ is a self-averaging quantity. Using the replica trick with a replica symmetric ansatz [15], one obtains the following result for the intensive entropy:

$$
\begin{aligned}
s &= \lim_{N\to\infty} \frac{\langle \ln\Omega\rangle}{N} \\
&= \mathop{\mathrm{extr}}_{\{q,\hat{q}\}} \left[-\frac{1}{2}(1-q)\hat{q} + \int_{-\infty}^{+\infty} Dz \, \ln\left(2\cosh(z\sqrt{\hat{q}})\right) \right. \\
&\quad \left. + \alpha \int_{-\infty}^{+\infty} Dt \, \ln\left(\int_{g(x\sqrt{1-q}-t\sqrt{q})>0} Dx \right) \right] \qquad (4)
\end{aligned}
$$

where $Dx = dx \, \exp(-x^2/2)/\sqrt{2\pi}$. The value of the parameter q, (with conjugated variable $\hat{q}$), which extremizes the expression in the right-hand side of this expression, has the physical meaning of the typical overlap $q = \vec{J_1}\cdot\vec{J_2}/N$ between two solutions $\vec{J_1}$ and $\vec{J_2}$ that correctly reproduce the classifications of the patterns under consideration. When this overlap goes to 1, one expects that the space of compatible solutions shrinks to zero so that this criterion could be used to identify the storage capacity. However, it is then found that the corresponding entropy is negative and that the predicted value of the storage capacity lies above the rigorous information theoretic upper bound $\alpha = 1$. One concludes that replica symmetry must be broken. To get around this difficulty, we will identify the storage capacity as the value of α for which the entropy becomes zero. The principle of combining replica symmetry with a zero entropy criterion has been shown to be equivalent to a one-step symmetry breaking in several cases [6,16–18], including the present problem for $K = 0$, and is widely believed [19,20] to give the correct result when the saddle point is locally stable. The resulting α versus K curve is reproduced in Fig. 1, together with the numerical results obtained from extensive simulations. The agreement between theory and simulation is very good. For $K = 0$, we of course recover the well-known $\alpha \approx 0.83$ value. Remarkably, the storage capacity increases with K until the maximal value $\alpha = 1$ is reached for $K = \sqrt{2\ln 2}$. For larger values of K, the storage capacity decreases to fall back, as expected, to the initial value $\alpha \approx 0.83$ for $K \to \infty$. The capacity $\alpha = 1$ for $K = \sqrt{2\ln 2}$ implies that the storage is optimal. A capacity $\alpha = 1$ was also found for a parity machine with $K \geq 2$ hidden units with nonoverlapping receptive fields [21].

We now investigate how the reversed wedge Ising perceptron can learn from examples generated by a teacher which is also a reversed wedge perceptron characterized by an Ising weight vector $\vec{T}$. As we mentioned before, the value $\alpha = 1$ is now a lower bound for reaching errorless generalization. We first present a simple annealed calculation, along the lines of the work by Gardner and Derrida [4,5]. It sheds light on the special properties of the reversed wedge for $K = \sqrt{2\ln 2}$, while it also provides an upper bound for the generalization threshold. Consider the reversed wedge Ising perceptrons $\vec{J}$ which have a given overlap $R = \vec{J}\cdot\vec{T}/N$ with the teacher. In the limit $N \to \infty$, the number of such perceptrons is given by

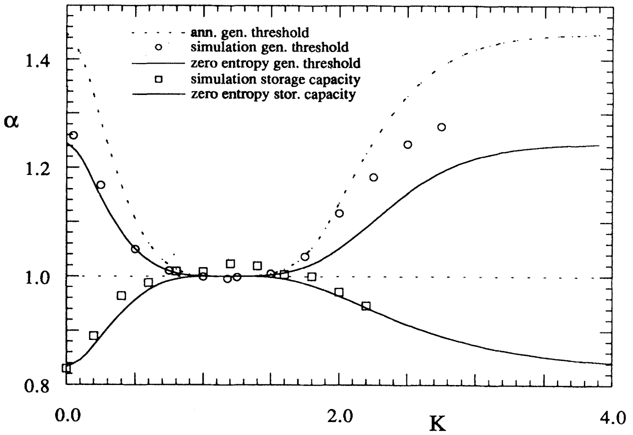

FIG. 1. Storage capacity [from the zero entropy criterion, cf. Eq. (4)] and threshold to errorless generalization [from the annealed calculation, cf. Eq. (7), and from the zero entropy criterion, cf. Eq. (8)] for the reversed wedge Ising perceptron, as a function of the width K of the wedge. The numerical results are an extrapolation from the results for $N = 13$ up to $N = 19$ (obtained through exact enumeration over the Ising $\vec{J}$ vectors and averaged over the choice of 10 000 pattern sets). For $K = \sqrt{2\ln 2}$ the quenched and annealed generalization errors coincide.

$$
\Omega_o(R) \sim \exp\left\{ N\left[-\frac{1-R}{2}\ln\frac{1-R}{2} - \frac{1+R}{2}\ln\frac{1+R}{2} \right] \right\} . \qquad (5)
$$

The corresponding generalization error $\epsilon(R)$, defined as the probability for disagreement between student and teacher on the classification of a randomly chosen pattern, is found to be

$$
\epsilon(R) = 2\int_{g(y)>0} Dy \int_{g(x\sqrt{1-R^2}+yR)<0} Dx \, . \qquad (6)
$$

For the particular choice $K = 0$, one recovers the familiar result $\epsilon(R) = \arccos R/\pi$. The average number of perceptrons with overlap R that is still compatible with the classification generated by the teacher on p random examples is given by $\langle \Omega(R)\rangle = \Omega_o(R)[1-\epsilon(R)]^p$. The compatible students, corresponding to the R value, that maximizes this expression, are on average exponentially more numerous than the others and determine the value of the annealed intensive entropy:

$$
\begin{aligned}
s_{\text{annealed}} &= \lim_{N\to\infty} \frac{\ln\langle\Omega\rangle}{N} \qquad (7) \\
&= \max_R \left[-\frac{1-R}{2}\ln\frac{1-R}{2} \right. \\
&\quad \left. - \frac{1+R}{2}\ln\frac{1+R}{2} + \alpha\ln[1-\epsilon(R)] \right] .
\end{aligned}
$$

The generalization error is obtained by substituting this

R value, which is a function of α, into Eq. (6). It is clear from Eq. (7) that the location of the maximum in R is determined by the competition between the number of perceptrons with a given value of R and their corresponding generalization error. $\Omega_o(R)$ is maximal for $R = 0$, which corresponds to the perceptrons orthogonal to the teacher. On the other hand, $\epsilon(R)$ is a monotonous decreasing function of R [with $\epsilon(-1) = 1$, $\epsilon(0) = 1/2$, and $\epsilon(1) = 0$], and the effect of the training examples is clearly to favor the student perceptrons that have a larger overlap with the teacher by preferentially eliminating those with large generalization error. As a result, one expects that the R value that maximizes the expression in the right-hand side of Eq. (7), will increase with α, starting with the value $R = 0$ (and $s_{\text{annealed}} = \ln 2$) for $\alpha = 0$, while the corresponding values of the intensive entropy and the generalization error decrease. Furthermore, the expression in the right-hand side of Eq. (7) is equal to zero for $R = 1$, so that the annealed entropy can never be negative. One thus finds that a discontinuous transition from finite generalization error ($R < 1$) to zero generalization error ($R = 1$) takes place at the critical threshold value of α for which s_{annealed} first becomes zero. This α value is reproduced in Fig. 1 as a function of K under the name of annealed approximation. For $K = 0$, one recovers the Gardner-Derrida result $\alpha \approx 1.45$ [4]. As the value of K increases, the critical value of α decreases until, again for $K = \sqrt{2\ln 2}$, the transition to perfect generalization occurs right at the lower bound $\alpha = 1$ imposed by information theory. Moreover, since the annealed calculation gives an upper bound, the latter result must also be exact. This surprising result can be understood from the fact that the generalization error $\epsilon(R)$ has a horizontal inflection point at $R = 0$, $\epsilon'(0) = \epsilon''(0) = 0$, for $K = \sqrt{2\ln 2}$. Consequently $R = 0$ is a maximum in the expression appearing in the right-hand side of Eq. (7) for all values of α. It remains the absolute maximum until the annealed entropy becomes zero. The picture that emerges is quite astonishing. For $\alpha < 1$, every new training example reduces the space of compatible perceptrons by half [$s_{\text{annealed}} = (1-\alpha)\ln 2$], while the students orthogonal to the teacher retain their exponentially dominant majority with corresponding generalization error $\epsilon(0) = 1/2$. At $\alpha = 1$, all the students except for the teacher himself have been eliminated, and a discontinuous transition takes place from no generalization $\epsilon(0) = 1/2$ to perfect generalization $\epsilon(1) = 0$, cf. Fig. 2.

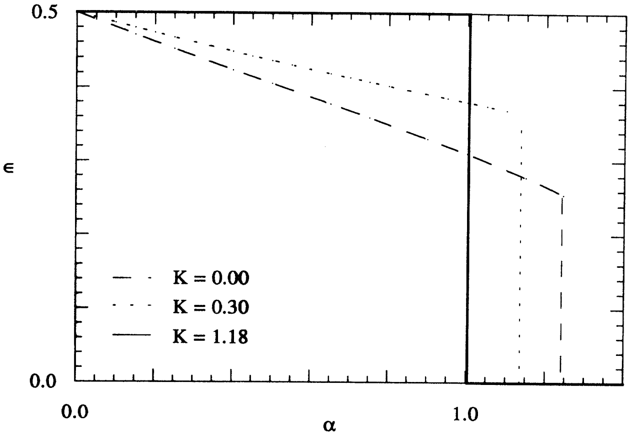

FIG. 2. Quenched generalization error of the reversed wedge Ising perceptron in function of α for several values of K. These curves are obtained by inserting the R value that extremizes Eq. (8) into Eq. (6).

Going beyond the annealed calculation, we have also performed a replica-symmetric calculation of the generalization error. The classification of the examples is no longer random, since it is generated by the perceptron teacher, and the intensive (quenched) entropy is now given by the following formula:

$$s = \operatorname*{extr}_{\{R,\hat{R}\}} \left[-\frac{1}{2}(1+R)\hat{R} + \int_{-\infty}^{+\infty} Dz \ln\left(2\cosh(z\sqrt{\hat{R}}+\hat{R})\right) \right.$$
$$\left. +2\alpha \int_{-\infty}^{+\infty} Dt \int_{g(y)>0} Dy \; \ln\left(\int_{g((x-t\sqrt{R})\sqrt{1-R}+yR)>0} Dx \right) \right] , \qquad (8)$$

where the R value that extremizes the right-hand side in Eq. (8) corresponds to the typical value of the overlap between a compatible student and the teacher perceptron. It determines, through Eq. (6), the generalization error in function of α. Furthermore, we identify the location of the discontinuous transition to perfect generalization as the value of α for which the entropy given in Eq. (8) becomes equal to zero. The zero entropy criterion combined with replica symmetry is again expected to give the correct result, although this has, to our knowledge, not yet been confirmed by an explicit replica symmetry-breaking calculation. The resulting α versus K curve is reproduced in Fig. 1. This curve lies, as required, under the curve corresponding to the annealed approximation, and is in good agreement with numerical simulations. In Fig. 2, we have also represented the generalization error $\epsilon(R)$ in function of α for some values of K, including the particular case $K = \sqrt{2\ln 2}$.

In conclusion, we have calculated the storage capacity and generalization error of a reversed wedge Ising perceptron, with transfer function defined by Eq. (1), and shown that the case $K = \sqrt{2\ln 2}$ saturates the information theoretic bounds for storage and generalization. Since this perceptron seems to store or utilize information in a more efficient way than the normal perceptron, while keeping its basic distributed and parallel architecture, we suggest that it may be an interesting choice as a building block for more complicated multilayer or multiconnected networks. At the same time, it illustrates that there are no inherent limitations, other than the information theoretic bounds, for such architectures to store or transfer information.

We acknowledge support from the Program on Inter-University Attraction Poles, Prime Minister's Office, Belgian Government, and from the NFWO Belgium (C.Vd.B). We also thank M. Bouten and J. Iwanski for helpful discussions and suggestions.

[1] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Santa Fe, 1991).

[2] H. S. Seung, H. Somolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[3] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. A **45**, 499 (1993).

[4] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).

[5] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).

[6] W. Krauth and M. Mézard, J. Phys. (France) **50**, 3057 (1989).

[7] W. Krauth and M. Opper, J. Phys. A **22**, L519 (1989).

[8] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).

[9] H. Sompolinsky, N. Tishby, and H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).

[10] B. Derrida, R. B. Griffiths, and A. Prügel-Bennett, J. Phys. A **24**, 4907 (1991).

[11] C. Van den Broeck and M. Bouten, Europhys. Lett. **22**, 223 (1993).

[12] T. Watkin and A. Rau, Phys. Rev. A **45**, 4102 (1992).

[13] K. Kobayashi, Network **2**, 237 (1991).

[14] G. Boffetta, R. Monasson, and R. Zecchina, J. Phys. A **26**, L507 (1993).

[15] E. Gardner, Europhys. Lett. **4**, 1205 (1987); J. Phys. A **21**, 257 (1988).

[16] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[17] D. J. Gross and M. Mézard, Nucl. Phys. **B240**, 431 (1984).

[18] J. Iwanski, Ph.D. thesis, Limburgs Universitair Centrum, 1994 (unpublished).

[19] M. Bouten, A. Komoda, and R. Serneels, J. Phys. A **23**, 2605 (1990).

[20] M. Gutfreund and Y. Stein, J. Phys. A **23**, 2613 (1990).

[21] E. Barkai and I. Kanter, Europhys. Lett. **14**, 107 (1991).