Learning in the hypercube: A stepping stone to the binary perceptron

M. Bouten, L. Reimers, and B. Van Rompaey
Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium
(Received 30 January 1998)

The learning problem for storing random patterns in a perceptron with binary weights can be facilitated by pretraining an appropriate precursor network with continuous weights. Unlike previous studies which compare the performance of different continuous-weight perceptrons on the hypersphere (spherical constraint), we also consider weight vectors constrained to the volume of the hypercube (cubical constraint). We compare the performance of the maximally stable networks on the hypersphere and in the hypercube, and show that the latter is superior for predicting the weights of the maximally stable binary perceptron. We further determine an upper bound for the fraction of binary weights that any precursor is able to predict correctly, and introduce a precursor in the hypercube that closely approaches this upper bound. We finally demonstrate the value of this hypercube precursor by carrying out simulations for a perceptron with up to 100 weights.
[S1063-651X(98)12808-8]

PACS number(s): 87.10.+e, 64.60.Cn

I. INTRODUCTION

In a recent paper [1], we introduced a learning strategy for the binary perceptron. It is based on previous work by Penney and Sherrington [2] which showed that a strong correlation exists between the signs of the synaptic weights of the continuous-weight perceptron of maximum stability (MSN) and those of the binary-weight perceptron of maximum stability (MSB). Since excellent algorithms [3,4] exist for determining the MSN weights, it is natural to try exploiting this correlation to collect valuable information about the MSB weights.

A first, albeit rather poor approximation of the MSB weight vector can be obtained by clipping all MSN weights. Penney and Sherrington [2] calculated that, near the saturation limit $\alpha = 0.83$, about 20% of these clipped weights differ in sign from the corresponding components of the MSB weight vector. To improve on the clipped weights, it is necessary to identify some of the incorrect components. On the basis of numerical experiments for small systems, Penney and Sherrington suggested that the components of the MSN, likely to give a wrong prediction by weight clipping, are to be found predominantly among the weakest MSN components. We have demonstrated [1] that this suggestion is indeed correct by focusing on the MSN weights that exceed a threshold value, and calculating the probability that they predict the correct sign for the MSB. Our result indicates that few errors will be generated by clipping the strongest 40% components of the MSN. However, the prediction, of the remaining MSB weights by clipping the weaker components of the MSN becomes increasingly more dubious. An additional learning stage therefore is necessary to determine these weights. Numerical simulations [1] for a perceptron with 50 input neurons confirm that such a two-stage learning procedure yields satisfactory agreement with theoretical expectations.

Although the MSN would seem like the obvious continuous-weight perceptron for serving as a precursor for the MSB—since both strive to maximize the stability—it turns out that the MSN is not the optimal choice. We have determined [1] the optimal continuous-weight perceptron that, on weight clipping, predicts the largest number of binary weights for the MSB correctly. We have, in addition, presented a simple cost function for use in numerical calculations that produces an excellent approximation to these optimal precursor weights.

In the present paper, we want to show how a significantly better continuous-weight precursor for the MSB can be constructed. To describe this precursor clearly, we first introduce our notation. As usual, we call N the number of input neurons of the perceptron, and $p = \alpha N$ the number of input vectors $\boldsymbol{\xi}^{\mu}$ $(\mu = 1, \ldots, p)$. These N-dimensional vectors are randomly chosen on the hypersphere $\boldsymbol{\xi} \cdot \boldsymbol{\xi} = N$. Without loss of generality, we can assume that all outputs are $+1$. The N weights of the binary perceptron are described by the weight vector $\boldsymbol{B}$ with components $B_i \in \{-1, +1\}$, $(i = 1, \ldots, N)$, while those of the continuous precursor perceptron are described by the weight vector $\boldsymbol{J}$ with components $J_i \in \mathbb{R}$, $(i = 1, \ldots, N)$. For the latter it is usual to impose the spherical constraint $\boldsymbol{J}^2 = N$. The input vectors $\boldsymbol{\xi}^{\mu}$ generate the fields $\Lambda_{\mu} = \boldsymbol{B} \cdot \boldsymbol{\xi}^{\mu} / \sqrt{N}$ in the binary perceptron, and $\lambda_{\mu} = \boldsymbol{J} \cdot \boldsymbol{\xi}^{\mu} / \sqrt{N}$ in the continuous perceptron. Learning the MSB involves finding the vector $\boldsymbol{B}$, such that $\Lambda_{\mu} \geqslant K_b$ $(\mu = 1, \ldots, p)$ with the largest possible value of the stability K_b. Similarly for the MSN, learning means finding the vector $\boldsymbol{J}$ such that $\lambda_{\mu} \geqslant K$ $(\mu = 1, \ldots, p)$ with the largest possible value of the stability K. More general learning rules for the continuous perceptron are usually formulated as an optimization problem [5–7]. Learning then consists of finding the vector $\boldsymbol{J}$ that minimizes a cost function of the general form $E(\boldsymbol{J}) = \Sigma_{\mu} V(\lambda_{\mu})$. The optimal continuous precursor weight vector referred to above corresponds to the optimal choice of the ''potential'' $V(\lambda)$ [1].

To construct an even better continuous precursor, one either has to modify the form of $E(\boldsymbol{J})$ or give up the spherical constraint. Perez Vicente, Carrabina, and Valderrana [8] used a modified cost function, containing a term $\Sigma_i (J_i^2 - 1)^2$ which shifts the minimum towards the binary vectors. Unfortunately, this new term also creates a huge number of

local minima in which the minimization becomes trapped. We keep the cost function unchanged, but replace the ''spherical'' constraint $\boldsymbol{J}^2=N$ by the ''cubical'' constraint $|J_i|\leqslant 1$ $(i=1,\ldots,N)$. The geometrical terminology is evident: the weight space in which learning has to proceed changes from the surface of a hypersphere to the volume of an inscribed hypercube. Two features of the hypercube make it attractive as a weight space for constructing a precursor for the MSB. Unlike the hypersphere, which has lost all information about the directions of the binary vectors, the hypercube retains a clear memory of them: they are the directions pointing toward the corners of the cube. Moreover, since these vectors are the longest vectors in the hypercube, they have an edge over the other vectors for generating large values of the fields λ_μ. This is nicely illustrated for the simple potential $V(\lambda)=-\lambda$, for which minimization in the hypercube directly leads to the clipped Hebb vector while minimization on the hypersphere yields the standard Hebb vector. Similarly, for more general potentials, as when we search for the maximally stable vector in the hypercube (MSC)—defined as the weight vector $\boldsymbol{J}$ with $|J_i|\leqslant 1$ $(i=1,\ldots,N)$ and satisfying $\lambda_\mu=\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}\geqslant K_c$ $(\mu=1,\ldots,p)$ with the largest possible value of the stability K_c—the binary vectors have a competitive advantage and become favored candidates. It is therefore reasonable to expect that the maximally stable vector in the hypercube will be a close neighbor of the MSB, closer than the maximally stable vector on the hypersphere. The second attractive feature of the hypercube is its convexity. If we want to use the MSC or any other learning rule in the hypercube as a precursor for the MSB, clearly a reliable learning algorithm is required to construct the precursor weights in the first place. Since the hypercube is a convex set, any cost function of the form $E(\boldsymbol{J})=\Sigma_\mu V(\lambda_\mu)$ with a convex potential $V(\lambda)$ will have a unique minimum which can easily be found by standard gradient descent algorithms.

In Sec. II, we study the optimization problem of a general cost function in the hypercube. In Sec. III, we examine the correlations between the weights of the MSB and those of different learning rules in the hypercube. More specifically, we calculate the probability that weight clipping produces the correct MSB weights. In Sec. IV, we focus on the strong components of the precursor, and demonstrate that they give a reliable prediction of the MSB weights. The quality of the hypercube precursor is further tested in Sec. V by carrying out numerical simulations for a perceptron with up to 100 input units. In Sec. VI, we discuss our results and look out to further improvements and applications of the hypercube precursor.

II. LEARNING IN THE HYPERCUBE

We consider an energy function of the form $E(\boldsymbol{J})=\Sigma_\mu V(\lambda_\mu)$, and want to determine the minimum of the energy in the hypercube $|J_i|\leqslant 1$ $(i=1,\ldots,N)$. We use the same definition of the fields $\lambda_\mu=\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ as on the hypersphere, even though the weight vector $\boldsymbol{J}$ in the hypercube is not normalized to N. This means that the fields no longer depend only on the angle between $\boldsymbol{J}$ and $\boldsymbol{\xi}^\mu$ but also on the length of $\boldsymbol{J}$. In the following, we will always assume that the potential $V(\lambda)$ is a convex function so that $E(\boldsymbol{J})$ has a unique minimum in the hypercube.

Following standard replica techniques [9], we calculate the free energy $f(\beta)$

$$-\beta f(\beta)=\operatorname*{Extr}_{q_0 q \hat{q}_0 \hat{q}}\left\{ q_0\hat{q}_0+\frac{1}{2}q\hat{q}+\int Dz\,\ln\left[\int_{-1}^{+1}dJ\,\exp\left(-\frac{2\hat{q}_0+\hat{q}}{2}J^2+zJ\sqrt{\hat{q}}\right)\right]\right.$$
$$\left.+\alpha\int Dz\,\ln\left[\int\frac{d\lambda}{\sqrt{2\pi(q_0-q)}}\exp\left(-\beta V(\lambda)-\frac{1}{2}\frac{(\lambda-\sqrt{q}z)^2}{q_0-q}\right)\right]\right\}, \tag{1}$$

where, as usual, $Dz=dz\,\exp[-z^2/2]/\sqrt{2\pi}$. The order parameters q_0 and q are defined as

$$q_0=\frac{1}{N}\sum_j J_j^a J_j^a, \quad q=\frac{1}{N}\sum_j J_j^a J_j^b, \tag{2}$$

with $\hat{q}_0$ and $\hat{q}$ as their conjugate variables. The labels a and b refer to different replicas, and replica symmetry has been assumed in deriving Eq. (1). This assumption is justified since $E(\boldsymbol{J})$ is supposed to have a single minimum in the hypercube.

To obtain the lowest possible value of the energy, we let $\beta\to+\infty$. Then $q\to q_0$ and $\hat{q}$ as well as $2\hat{q}_0+\hat{q}$ tend to infinity. We therefore introduce three new parameters to replace q, $\hat{q}_0$, and $\hat{q}$:

$$x=\beta(q_0-q), \quad y=(2\hat{q}_0+\hat{q})(q_0-q), \quad s=\frac{\sqrt{\hat{q}}}{2\hat{q}_0+\hat{q}}. \tag{3}$$

The lowest energy e_0 is then obtained as

$$e_0=-\operatorname*{Extr}_{q_0 x y s}\left\{\frac{y}{2x}\left[q_0-ys^2-\int Dz\,\min_{J\in[-1,+1]}(J^2-2zsJ)\right]\right.$$
$$\left.-\alpha\int Dz\,\min_{\lambda}\left[V(\lambda)+\frac{(\lambda-\sqrt{q_0}z)^2}{2x}\right]\right\} \tag{4}$$

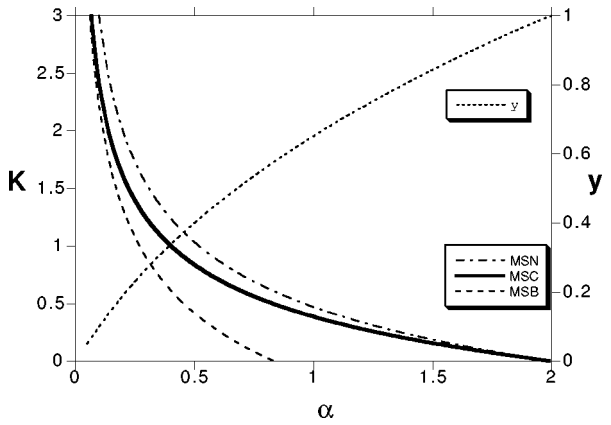

FIG. 1. Maximum stability K obtained for three different types of constraints on the weights: binary, spherical, and hypercube. The rescaled hypercube stability $K_c(\alpha)/\sqrt{q_0}$ is shown to obtain a meaningful comparison with the other cases. The dotted curve shows $y(\alpha)$, the fraction of weights in the maximally stable hypercube vector that have magnitude smaller than 1.

The extremum yields four saddle-point equations which determine the four order parameters q_0, x, y, and s as functions of α. These equations are written down in Appendix A. In comparison with the corresponding minimization of $E(\boldsymbol{J})$ on the hypersphere $\boldsymbol{J}^2=N$, where only the parameter x appears, we need three extra order parameters in the hypercube. The meaning of q_0N follows from Eq. (2) as the norm of the lowest-energy vector $\boldsymbol{J}$. The meaning of y and s becomes clear when we write down the distribution of the components of $\boldsymbol{J}$:

$$P(J)=\frac{1}{\sqrt{2\pi}s}e^{-(J^2/2s^2)}\theta[1-|J|] + H\left[\frac{1}{s}\right](\delta[J-1]+\delta[J+1]) \tag{5}$$

where, as usual, $H[u]=\int_u^\infty Dz=\mathrm{Erfc}[u/\sqrt{2}]/2$. The components of $\boldsymbol{J}$ that have a magnitude smaller than 1 follow a Gaussian distribution with variance s^2. The two tails of this Gaussian are compressed into two δ peaks at -1 and $+1$. The saddle-point equation (A1) shows that $1-y$ represents the fraction of components of J with magnitude equal to 1. Note that the form of the distribution $P(J)$ changes with α through its dependence on the parameter s. This entails that the fraction of components greater than a fixed value J_0, given by $\int_{J_0}^\infty P(J)dJ=H[J_0/s]$, also varies with α.

We now specialize to the perceptron of maximum stability in the hypercube. Results for the MSC are obtained from the general expression (4) by choosing the potential $V(\lambda)=(K_c-\lambda)^2\theta(K_c-\lambda)$, and assigning the value infinity to the order parameter x [6,9]. The four saddle-point equations now determine the three remaining order parameters q_0, s, and y, as well as the stability parameter K_c, as functions of α. Figure 1 shows the solution for $K_c(\alpha)$ and $y(\alpha)$. To obtain a meaningful comparison with the value of the maximum stabilities $K(\alpha)$ for the MSN and $K_b(\alpha)$ for the MSB, we plot the ''normalized'' value $K_c(\alpha)/\sqrt{q_0}$, correcting for the shorter length of the MSC weight vector. Not surprisingly, when $\alpha\to 2$, $K_c/\sqrt{q_0}$ goes to zero like $K(\alpha)$ for the MSN. More interesting, however, is the behavior at the other end $\alpha\to 0$, where $K_c/\sqrt{q_0}$ is found to coincide with $K_b(\alpha)$. The overall impression emerging from Fig. 1 is that the MSC interpolates smoothly between the MSB (at $\alpha=0$) and the MSN (at $\alpha=2$). This impression is substantiated by looking at $y(\alpha)$, the fraction of MSC components that have a magnitude smaller than 1. This fraction increases from 0 at $\alpha=0$—indicating that the MSC at $\alpha=0$ is indeed a binary vector—up to the value 1 at $\alpha=2$, indicating that $P(J)$ transforms into the pure Gaussian distribution of the MSN. At the saturation limit $\alpha=0.83$ where the MSB ceases to exist, 43% of all MSC weights are still binary. These general findings strengthen our confidence in the MSC weight vector as an excellent precursor for the MSB, especially at small values of α, but gradually declining in quality when α increases.

III. PRECURSORS FOR THE BINARY PERCEPTRON

In this section we estimate the significance of the MSC and other hypercube vectors as precursors of the MSB. As preparation for the subsequent discussion, we first recall some general characteristics of the MSB which are derived from replica calculations in the thermodynamic limit [2,10]. Unlike the continuous MSN and MSC weight vectors which are unique vectors for any value of α, different binary vectors exist with the same maximum value $K_b(\alpha)$ of the stability. The different vectors of the MSB ensemble have a typical mutual overlap Q which decreases from 1 at $\alpha=0$ down to 0.56 at $\alpha=0.83$. Since it is impossible to distinguish the individual vectors, all theoretical results relate to averages over this ensemble of MSB vectors. The implication is that any algorithm for constructing the MSB on the basis of theoretical arguments will at best be directed toward the average $\langle\boldsymbol{B}\rangle$ of this ensemble of vectors, not to a particular individual vector $\boldsymbol{B}$. The lack of uniqueness of the MSB weights constitutes a major obstacle to any theoretical algorithm.

An obvious measure for gauging the quality of a continuous precursor vector $\boldsymbol{J}$ is given by the proportion of binary weights $\boldsymbol{J}$ is able to predict correctly. This number can be derived from the joint probability distribution $P(B,J)$ of corresponding components in the weight vectors $\boldsymbol{B}$ and $\boldsymbol{J}$. To calculate $P(B,J)$, we follow the approach of Wong, Rau, and Sherrington [11], and consider the combined system of a binary and a continuous perceptron, both trained by the same random input vectors. The weight vector of the continuous perceptron is defined by an energy function $E(\boldsymbol{J})$ in the hypercube, while the weight vector of the binary perceptron is the MSB. Besides the order parameters of the separate perceptrons, two new order parameters appear that relate to both perceptrons: the overlap r of the continuous vector $\boldsymbol{J}$ with the average $\langle\boldsymbol{B}\rangle$ of the binary vectors, and its conjugate parameter $\hat{r}$. The saddle-point equations for these new order parameters are written down in Appendix B. There, as well as in all subsequent equations, r and $\hat{r}$ generally appear in combination with an order parameter from each separate perceptron. It is expedient to introduce, a new notations for these combinations:

$$\gamma = \frac{r}{\sqrt{qQ}}, \quad \hat{\gamma} = \frac{\hat{r}}{\sqrt{\hat{q}\hat{Q}}}. \tag{6}$$

The parameters q and $\hat{q}$ are order parameters of the continuous perceptron $\boldsymbol{J}$ encountered in Sec. II. The parameter Q is the mutual overlap of two binary vectors in the MSB ensemble discussed above, and $\hat{Q}$ is its conjugate [10]. The new parameters γ and $\hat{\gamma}$ have some further advantage over r and $\hat{r}$. The denominator $\sqrt{qQ}$ corrects for the length of both $\boldsymbol{J}$ and $\langle \boldsymbol{B} \rangle$, so that γ equals the cosine of the angle between $\boldsymbol{J}$ and $\langle \boldsymbol{B} \rangle$. Also, while both $\hat{r}$ and $\hat{q}$ tend to infinity, $\hat{\gamma}$ retains a finite value. Letting $\hat{\gamma}$ increase and tend to its maximum value 1 moves $\boldsymbol{J}$ closer to $\langle \boldsymbol{B} \rangle$, so γ also tends to its maximum value. The parameter $\hat{\gamma}$ will play an important role in the following discussion.

The probability distribution $P(B,J)$ can be expressed as

$$P(B,J) = \int \int D_{\hat{\gamma}}(u,v) \frac{1}{2}[1 + \tanh(B\sqrt{\hat{Q}}v)] \times \frac{\exp\left(-\frac{2\hat{q}_0 + \hat{q}}{2}(J - su)^2 \right)}{\int_{-1}^{+1} dj \, \exp\left(-\frac{2\hat{q}_0 + \hat{q}}{2}(j - sv)^2 \right)}. \tag{7}$$

We recall that $B \in \{-1,+1\}$ and $J \in [-1,+1]$. The shorthand notation $D_{\hat{\gamma}}(u,v)$ stands for the two-dimensional Gaussian measure with correlation $\hat{\gamma}$

$$D_{\hat{\gamma}}(u,v) = \frac{du\,dv}{2\pi\sqrt{1-\hat{\gamma}^2}} \times \exp\left(-\frac{1}{2(1-\hat{\gamma}^2)} [u^2 + v^2 - 2\hat{\gamma}uv] \right). \tag{8}$$

The integrand in Eq. (7) is the product of two factors, each factor relating to one of the two perceptrons only. The first factor relates to the binary vector $\boldsymbol{B}$ via the order parameter $\hat{Q}$ of the MSB. The second factor relates to the hypercube vector $\boldsymbol{J}$ via the order parameter s encountered in Sec. II. Recall that the other combination of parameters $2\hat{q}_0 + \hat{q}$ appearing in Eq. (7) tends to infinity. The second factor in Eq. (7) therefore has the character of a δ function.

Due to the symmetry $P(B,J) = P(-B,-J)$, we can confine the following argument to the value $B = +1$ only. The fraction of positive components of $\boldsymbol{J}$ that correctly predict the binary component $B = +1$ is given by

$$f(\alpha) = \frac{\int_0^\infty P(1,J)\,dJ}{\int_0^\infty P(J)\,dJ}. \tag{9}$$

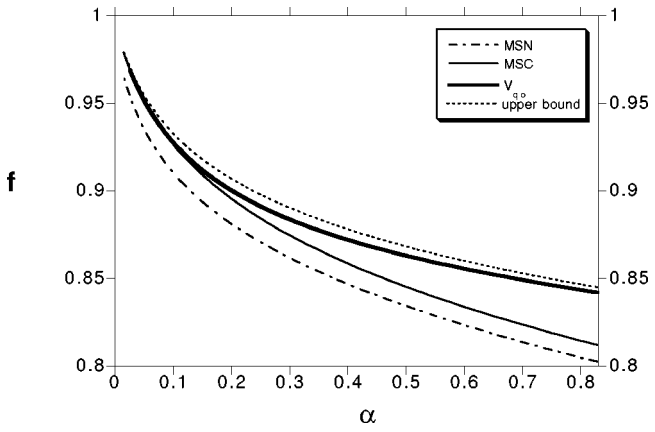

FIG. 2. Fraction of binary weights in the MSB that are correctly predicted by clipping all continuous weights of the MSN, the MSC, and the quasioptimal hypercube precursor V_{qo}. The dotted line shows the upper bound (11).

We have explicitly indicated that f depends on α which enters via the order parameters in $P(B,J)$. Straightforward calculation of the integral yields

$$f(\alpha) = \frac{1}{2} + \int_{-\infty}^{+\infty} Du \, \tanh(\sqrt{\hat{Q}}u) H\left(\frac{-\hat{\gamma}u}{\sqrt{1-\hat{\gamma}^2}} \right). \tag{10}$$

The whole dependence on the choice of potential $V(\lambda)$ in the cost function enters via the parameter $\hat{\gamma}$. It is easy to see that the value of the integral grows with $\hat{\gamma}$. Hence a sharp upper bound for the fraction $f(\alpha)$ can be obtained by taking the limit $\hat{\gamma} \to 1$:

$$f(\alpha) \leqslant \frac{1}{2} + \int_0^\infty Du \, \tanh(\sqrt{\hat{Q}}u). \tag{11}$$

This bound only depends on the conjugate parameter $\hat{Q}$ of the binary perceptron, and consequently cannot be surpassed by any choice of potential $V(\lambda)$ in the hypercube. Because the value of $\hat{Q}$ is finite for all $\alpha > 0$, the upper bound is less than 1, and decreases steadily with growing α. The finite value of $\hat{Q}$ is connected with the lack of uniqueness of the MSB weights, as reflected by the mutual overlap Q being smaller than 1. Since for a perceptron with a spherical constraint, exactly the same expression (10) was obtained in Ref. [1], the upper bound (11) is valid for any potential on the hypersphere as well.

Figure 2 shows the fraction (10) of binary components of the MSB, correctly predicted by clipping all weights of the MSC. For comparison, we also show the corresponding fraction predicted by clipping the weights of the MSN [2] as well as the upper bound (11). As expected, for small values of α, the MSC achieves a substantial improvement, and approaches the upper bound very closely. At larger values of α, the improvement is smaller and the separation from the upper bound remains considerable. In an attempt to bridge the gap, we have selected a different precursor vector in the hypercube using the potential

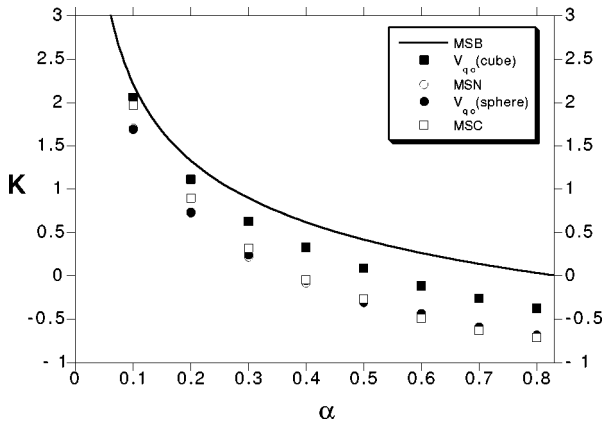

FIG. 3. Minimum pattern stability K for a perceptron with $N=50$ as determined by numerical simulations. The binary weights are obtained through full clipping from four different precursors. At small α, the results from the MSN coincide with those from the spherical quasioptimal precursor and, at large α, with the MSC results.

$$V_{qo}(\lambda)=\begin{cases}1/(\lambda-K_b) & \text{when } \lambda>K_b\\ +\infty & \text{when } \lambda<K_b\end{cases}. \tag{12}$$

This simple potential, which we call the ''quasioptimal potential,'' was introduced in Ref. [1] as a substitute for the optimal potential on the hypersphere. The strong repulsion, away from the boundaries of the Gardner volume [12] with stability K_b, pushes the minimizing vector toward its center of mass [7]. The center of mass would be the optimal precursor, given that the sole information available about the position of the MSB is that it lies on the boundary of the Gardner volume with stability K_b [13]. The resulting value of $f(\alpha)$ is also shown in Fig. 2. It achieves a remarkably large improvement at large values of α, and closely approaches the upper bound (11) over the whole interval.

Numerical simulations confirm the superiority of the hypercube precursor that minimizes $E(\boldsymbol{J})$ with the quasioptimal potential (12). Figure 3 shows the minimum stability $K(\alpha)$ of the binary vector obtained by clipping all weights of this hypercube precursor for a perceptron with 50 input neurons. For comparison, we also show the minimum stability obtained by clipping three other precursors: the MSC, the MSN, and the spherical precursor that minimizes the cost function $E(\boldsymbol{J})$ with potential (12). The quasioptimal hypercube precursor stands out well above the results of the other precursors over the whole range of α. It narrows the gap between the MSN and the theoretical curve $K_b(\alpha)$ by more than half. The outcome from the other two precursors is intermediate. As expected, the MSC result lies close to the quasioptimal hypercube precursor at small α, but rapidly deteriorates when α increases to coincide with the MSN result at large α. The outcome from the quasioptimal spherical precursor coincides with the MSN result at small α, and moves only slightly above the MSN at large α. These simulations confirm the superiority of the quasioptimal hypercube precursor, stressing that both the hypercube constraint and the quasioptimal potential are essential for excellent performance.

IV. RELIABLE COMPONENTS OF THE HYPERCUBE PRECURSORS

Despite this large improvement, even the best of all possible precursors in the hypercube—represented by the upper bound (11)—fails to predict about 16% of the MSB components correctly near $\alpha=0.83$. A further learning stage therefore will always be necessary in which the incorrect components have to be identified and corrected. Again we suspect the weak components of the hypercube vector $\boldsymbol{J}$ to be the dubious ones, while we expect the stronger components to be more reliable. To check this expectation, we focus on the components of $\boldsymbol{J}$ that are greater than a threshold value $J_0>0$. The fraction of these components that, on weight clipping, correctly predict the corresponding component of $\boldsymbol{B}$ is given by

$$f_{J_0}(\alpha)=\frac{\int_{J_0}^{\infty}P(1,J)\,dJ}{\int_{J_0}^{\infty}P(J)\,dJ}. \tag{13}$$

Straightforward calculation of the integral yields

$$f_{J_0}(\alpha)=\frac{1}{2}+\frac{1}{2H\left(\frac{J_0}{s}\right)}\int Du\,\tanh(\sqrt{\hat{Q}}u)H\left(\frac{\frac{J_0}{s}-\hat{\gamma}u}{\sqrt{1-\hat{\gamma}^2}}\right), \tag{14}$$

which, for $J_0=0$, returns to expression (10) for $f(\alpha)$. The value of the integral again grows steadily with $\hat{\gamma}$, so that an upper bound for $f_{J_0}(\alpha)$ can be obtained by taking the limit $\hat{\gamma}\to 1$:

$$f_{J_0}(\alpha)\leqslant\frac{1}{2}+\frac{1}{2H\left(\frac{J_0}{s}\right)}\int_{J_0/s}^{\infty}Du\,\tanh(\sqrt{\hat{Q}}u). \tag{15}$$

At this stage, it is important to recall that the components of the hypercube precursor of magnitude greater than J_0 form a fraction $2H[J_0/s]$ of the total number of components. For a fixed value of J_0, this fraction changes with α because the order parameter s changes with α. In a like manner, this fraction also changes for different choices of the potential $V(\lambda)$. If we want to compare the value of $f_{J_0}(\alpha)$ for different precursors, it would therefore not be reasonable to fix the value of J_0, because different numbers of components would be compared for different precursors. For a meaningful comparison, in which the same number of precursor components are examined, we have to fix the value of the ratio J_0/s.

In Fig. 4, we plot the fraction $f_{J_0}(\alpha)$ for three values of J_0/s, corresponding to the 40%, 60% and 80% strongest components of the MSN, the MSC, and the quasioptimal hypercube precursor. For each value of J_0/s, the upper bound (15) is also shown. It forms a standard against which the performance of the different precursors can be measured. The figure demonstrates the manifest superiority of the hy-

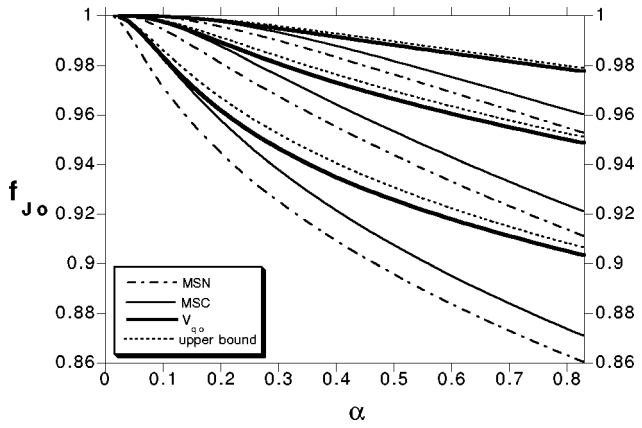

FIG. 4. Fraction of binary weights in the MSB that are correctly predicted by clipping the strongest 40% (top), 60% (middle), or 80% (bottom) components of the MSN, the MSC, or the quasioptimal hypercube precursor V_{qo}. The dotted curve shows the upper bound (15).

percube precursor with the quasioptimal potential. Compared to the MSN result, there is a substantial improvement for all values of α. For small α, this is completely due to the hypercube constraint, since it is also obtained for the MSC precursor. At larger α, part of the improvement comes from the hypercube constraint, but the major part comes from the quasioptimal potential. The numerical value of $f_{J_0}(\alpha)$ indicates that the 40% strongest components of the quasioptimal hypercube precursor are highly reliable predictors of the MSB weights. The probability of making a wrong prediction is small when the 60% strongest components are clipped. It increases further for the next 20% components. Comparison with Fig. 2, however, indicates that the greatest concentration of incorrect predictions occurs among the 20% weakest components of $\boldsymbol{J}$.

V. NUMERICAL SIMULATIONS

So far, we have focused on the *number* of binary weights that are correctly predicted by clipping various fractions of components of the continuous precursors. For practical purposes, a more appropriate quality measure of the precursor is the maximum possible value of the *stability* that can be attained, after clipping various fractions of strong components, by a perfect learning procedure for the remaining binary weights. As it is difficult to determine this maximum stability analytically, we rely on numerical simulations to acquire the relevant information. Clearly, determining the maximum possible value of the stability precludes any approximation in the determination of the remaining binary weights. This implies that the full enumeration method [14,15] has to be applied to obtain these weights.

The numerical simulations were carried out using the following simple ''learning algorithm.'' We start by minimizing the cost function $E(\boldsymbol{J})$ in the hypercube to determine the quasioptimal precursor $\boldsymbol{J}$. This is a fast and straightforward calculation, because $E(\boldsymbol{J})$ has a single minimum in the hypercube. In the second step, we clip a fraction of the strong components of $\boldsymbol{J}$, assuming that they can be trusted to provide an excellent prediction for the MSB weights. The primary objective of our simulations is to verify this assumption. In the last step, we determine the remaining weights by the full enumeration method. Since enumerating more than 25 weights becomes time consuming, the number of weights left over after clipping may not exceed 25.

In the hypercube precursor $\boldsymbol{J}$, many components have magnitude 1, especially when α is small. In cases when more components have magnitude 1 than we intend to clip, the question arises as to how the strongest ones are to be identified. We tackle this problem by adding to the cost function $E(\boldsymbol{J})$ a suitable ''perturbation'' which partly lifts the degeneracy of the components of magnitude 1. An obvious choice of perturbation is a term $\rho \boldsymbol{J}^2$ with $\rho > 0$. This term clearly exerts a force that pulls the minimum of the cost function toward the origin $\boldsymbol{J}=0$. Fine tuning ρ makes it possible to reduce the fraction $1-y$ of magnitude 1 components to a prescribed value. The required strength of ρ can simply be calculated by adding the term $\rho \boldsymbol{J}^2$ to the cost function $E(\boldsymbol{J})$. This produces an additional term ρq_0 in expression (4) of the lowest energy e_0. Since this term depends solely on the order parameter q_0, only the saddle-point equation (A4) will be altered, an extra term $-2x\rho$ being added to the left hand side of this equation. For given α, we are now free to choose the fraction $1-y$ of components that have magnitude 1 (smaller than the value obtained when $\rho=0$). The four saddle-point equations then determine the parameters q_0, s, x, and ρ.

It is to be noted that the additonal term $\rho \boldsymbol{J}^2$ does not affect the convexity of the cost function (when $\rho > 0$), so that a unique minimum continues to exist in the hypercube.

Figures 5 and 6 show results from our numerical simulations. The minimum pattern stability K is plotted as a function of α for the best binary vector obtained from the quasioptimal precursor in the hypercube. Each data point represents the average over 200 samples. The input vectors used in the simulation are random Gaussian patterns [16]. Figure 5 shows results for the relatively small system $N=40$. Figure 5(a) shows the effect of clipping different fractions of the quasioptimal hypercube precursor. When only 16, i.e., 40%, of the precursor components are clipped, our theoretical curves in Fig. 4 predict that all clipped components are very likely to give the correct binary weight. The simulations beautifully confirm this prediction, the numerical value of the stability lying even above the theoretical curve $K_b(\alpha)$ for all values of α. When 24, i.e., 60%, of the components are clipped, we deduce from Fig. 4 that, at large values of α, at least one of the clipped components is likely to produce an incorrect binary weight. The numerical simulations continue to display excellent agreement with the theoretical curve for all values of α, but the 60% clipping results lie very slightly below the 40% clipping points at large values of α. Figure 5(b) compares results from two different precursors: the quasioptimal hypercube and the MSN precursor. In both cases, the strongest 60% components were clipped. Although a doubling of the number of incorrect binary weights is to be expected for the MSN, the numerical results continue to agree nicely with $K_b(\alpha)$ at small α, but the fit deteriorates slightly for large values of α.

Figure 6 shows results for larger networks $N=75$ and 100. In these cases, a much larger number of precursor components have to be clipped because our computational capabilities restrict enumeration to 25 components. This entails

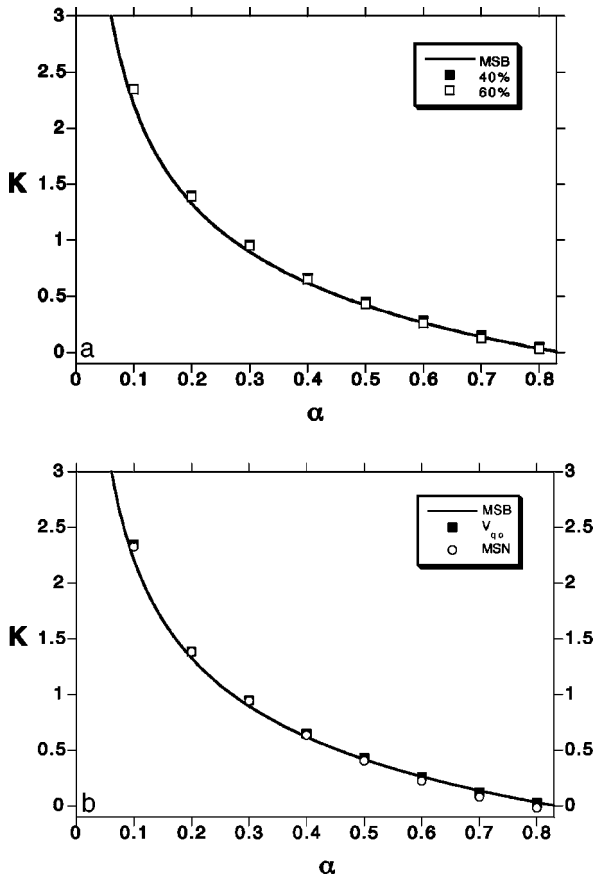

FIG. 5. Minimum pattern stability K for a perceptron with $N=40$. (a) shows the value of K when 40% or 60% of the quasioptimal hypercube precursor components are clipped. (b) compares the MSN and the quasioptimal hypercube precursor when the strongest 60% components are clipped. The full line shows the theoretical curve $K_b(\alpha)$.

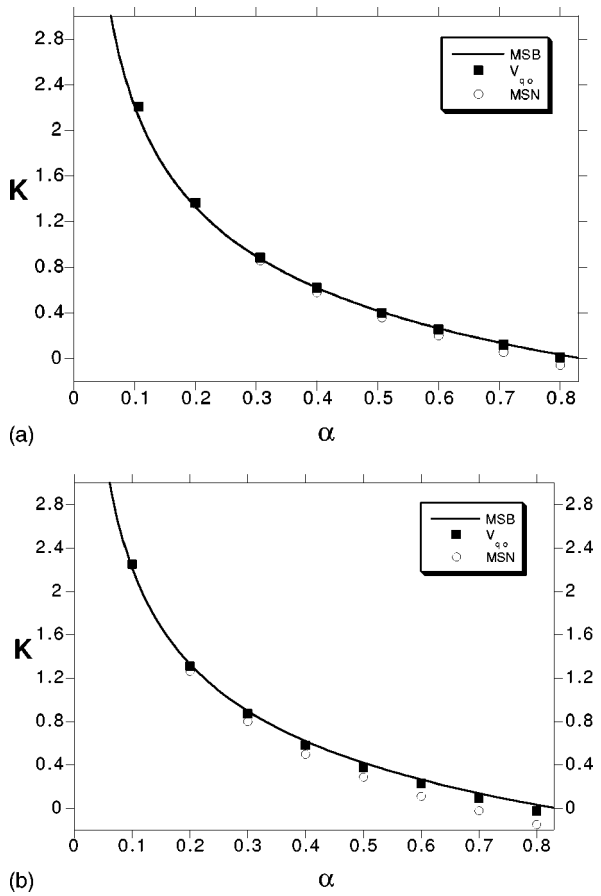

FIG. 6. Minimum pattern stability K for a perceptron with $N=75$ (a) and $N=100$ (b). Both figures compare the MSN and the quasioptimal hypercube precursor. Since only 25 components are enumerated, a considerable number of incorrect clipped weights are expected. The agreement with the theoretical curve $K_b(\alpha)$ nevertheless remains very satisfactory. The full line shows the theoretical curve $K_b(\alpha)$.

that many more clipped components will predict an incorrect binary weight. For the quasioptimal hypercube precursor, the number of incorrect predictions, as deduced from Fig. 4, can be estimated as equal to three for $N=75$ and equal to six for $N=100$ at large α. The interesting point now is to investigate how these many incorrect weights effect the value of the minimum stability K. Surprisingly, for $N=75$ with 50 components of the hypercube precursor clipped, the numerical value obtained from the simulations still follows nicely the theoretical curve $K_b(\alpha)$ over the whole range of α. For $N=100$ with 75 clipped components, the agreement with $K_b(\alpha)$ remains excellent for small α and the deviation at large values of α is small. This unexpected result indicates that the various incorrect weights generated by clipping the hypercube precursor do not destroy the high stability, as they might have done, but only affect a small reduction of its value. Apparently, the value of these particular weights is not crucial for obtaining a large value for the minimum stability. This is a very gratifying result, because it shows that the hypercube precursor performs even better than could be expected from Fig. 4. For comparison, we again show the corresponding results for the MSN precursor in which case an even larger number of incorrect binary weights are predicted. The MSN, however, also performs splendidly at small α, but the results become markedly less good at large values of α.

VI. DISCUSSION

In this paper, we have examined continuous-weight vectors in the hypercube as precursors for learning the binary weigths of the MSB. We have demonstrated that the vector $\boldsymbol{J}$ that minimizes the cost function $E(\boldsymbol{J})=\Sigma_\mu V(\lambda_\mu)$ with the potential (12) is nearly optimal in its ability to predict the largest number of MSB weights correctly. We have shown, in addition, that the strongest components of $\boldsymbol{J}$ are highly reliable predictors of the binary weights while the majority of uncertain predictors are to be found among the weakest components of $\boldsymbol{J}$. The substantial increase in predictive power of our new precursor, in comparison to previous precursors like the MSN, is achieved through both the hypercube constraint and the quasioptimal potential.

The analytical results as well as numerical simulations indicate that the hypercube precursor can play a very helpful role in reducing the overall difficulty of the learning problem for the binary perceptron. For small values of α, at least 60% of the binary weights can be reliably obtained from the precursor, while for larger values of α, still 40% of the binary weights are correctly predicted. This replaces the original learning problem by a simpler one of smaller size. Our numerical simulations for $N=75$ and 100 moreover indicate

that even considerably larger fractions of the hypercube precursor components may be clipped with only a tiny reduction in the value of the minimum stability as a result. This numerical finding suggests that the hypercube precursor correctly predicts all the binary weights that are essential for obtaining a high value of the minimum stability, and that those components where the precursor fails to predict the correct sign are not crucial for a high stability. In our simulations, we have used the full enumeration method to learn the weights of the reduced problem, restricting for computational reasons the number of weights to 25. More intelligent methods, like branch and bound [17] could be applied to enlarge this number up to 40.

The hypercube precursor is likely to play a similar simplifying role in other learning problems with discrete weights. We are currently exploring its usefulness in the storage problem for the diluted binary perceptron [18] as well as in supervised learning with a binary teacher [19].

ACKNOWLEDGMENTS

We thank K. Y. M. Wong for meaningful comments on the manuscript. We thank the Inter-University Attraction Poles of the Belgian Government for financial support. B.V.R. also acknowledges support from the FWO, Belgium.

APPENDIX A

The extremum of e_0 in Eq. (4) leads to the following four saddle-point equations for the parameters y, s, x, and q_0:

$$y = 1 - 2H\left(\frac{1}{s}\right), \tag{A1}$$

$$q_0 = s^2 y + 1 - y - \sqrt{\frac{2}{\pi}}\, s e^{-1/2s^2}, \tag{A2}$$

$$s^2 y^2 = \alpha \int D\tau\, [\lambda_0(\sqrt{q_0}\tau, x) - \sqrt{q_0}\tau]^2, \tag{A3}$$

$$y = -\frac{\alpha}{\sqrt{q_0}} \int \tau D\tau\, [\lambda_0(\sqrt{q_0}\tau, x) - \sqrt{q_0}\tau]. \tag{A4}$$

The first two equations are independent of the choice of the potential $V(\lambda)$. The last two equations depend on $V(\lambda)$ via the function $\lambda_0(z,x)$ defined as

$$\lambda_0(z,x) = \mathrm{Arg}\ \min_{\lambda}\left[V(\lambda) + \frac{(\lambda - z)^2}{2x}\right]. \tag{A5}$$

APPENDIX B

For the combined system of a binary and a continuous perceptron considered in Sec. III, the following equations are obtained for the overlap r and its conjugate $\hat{r}$:

$$r = \int\int D_{\hat{\gamma}}(u,v) \tanh[\sqrt{\hat{Q}}u] J_{\min}(sv), \tag{B1}$$

$$\hat{r} = \alpha \int\int D_{\gamma}(\sigma,\tau)\, \frac{1}{\sqrt{2\pi(1-Q)}}\, \frac{\exp\left(-\dfrac{(K_b - \sqrt{Q}\sigma)^2}{2(1-Q)}\right)}{H\left[\dfrac{K_b - \sqrt{Q}\sigma}{\sqrt{1-Q}}\right]} \times \left(\frac{\lambda_0(\sqrt{q_0}\tau, x) - \sqrt{q_0}\tau}{q_0 - q}\right), \tag{B2}$$

where γ and $\hat{\gamma}$ are defined in Eq. (6), $D_{\hat{\gamma}}(u,v)$ is the Gaussian measure (8), and $J_{\min}(z)$ is defined by

$$J_{\min}(z) = \mathrm{Arg} \min_{J \in [-1,+1]} (J^2 - 2zJ). \tag{B3}$$

Equation (B1) simply expresses that r is the average value $\langle BJ \rangle = \Sigma_B \int dJ P(B,J) BJ$ of the product of corresponding components of $\boldsymbol{J}$ and $\boldsymbol{B}$. It does not explicitly depend on $V(\lambda)$. The form of Eq. (B2), on the other hand, does depend on the potential $V(\lambda)$ via the function $\lambda_0(z,x)$ defined in Eq. (A5). When $q \to q_0$ it is clear that $\hat{r}$ tends to infinity. But $\hat{\gamma} = \hat{r}/\sqrt{\hat{q}\hat{Q}}$ is finite, since $\sqrt{\hat{q}}(q_0 - q)$ remains finite.

[1] L. Reimers, M. Bouten, and B. Van Rompaey, J. Phys. A **29**, 6247 (1996).
[2] R. Penney and D. Sherrington, J. Phys. A **26**, 6173 (1993).
[3] J. K. Anlauf and M. Biehl, Europhys. Lett. **10**, 687 (1989).
[4] W. Krauth and M. Mézard, J. Phys. A **20**, L745 (1987).
[5] K. Y. M. Wong and D. Sherrington, J. Phys. A **23**, 4659 (1990).
[6] M. Griniasti and H. Gutfreund, J. Phys. A **24**, 715 (1991).
[7] M. Bouten, J. Schietse, and C. Van den Broeck, Phys. Rev. E **25**, 1958 (1995).
[8] C. J. Perez Vicente, J. Carrabina, and E. Valderrana, Network **3**, 165 (1992).
[9] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[10] W. Krauth and M. Mézard, J. Phys. (France) **50**, 3057 (1989).
[11] K. Y. M. Wong, A. Rau, and D. Sherrington, Europhys. Lett. **19**, 559 (1992).
[12] E. J. Gardner, J. Phys. A **21**, 257 (1988).
[13] T. L. H. Watkin, Europhys. Lett. **21**, 871 (1993).
[14] B. Derrida, R. B. Griffiths, and A. Prügel-Bennett, J. Phys. A **21**, 4907 (1991).
[15] W. Nadler and W. Fink, Phys. Rev. Lett. **78**, 555 (1997).
[16] W. Krauth and M. Opper, J. Phys. A **22**, L519 (1989).
[17] G. Milde and S. Kobe, J. Phys. A **30**, 2349 (1997).
[18] J. Iwanski, J. Schietse, and M. Bouten, Phys. Rev. E **52**, 888 (1995).
[19] J. Schietse, M. Bouten, and C. Van den Broeck, Europhys. Lett. **32**, 279 (1995).