Selecting examples for perceptrons

1992 J. Phys. A: Math. Gen. 25 113

(http://iopscience.iop.org/0305-4470/25/1/016)

J. Phys. A: Math. Gen. 25 (1992) 113–121. Printed in the UK

Selecting examples for perceptrons

T L H Watkin and A Rau

Department of Theoretical Physics, University of Oxford, 1 Keble Rd, Oxford OX1 3NP, UK

Received 4 February 1991, in final form 27 June 1991

Abstract. A technique is introduced for analysing the learning process of perceptrons which continually select their own examples (the most efficient training algorithm yet devised). We predict a 42% reduction in the number of examples 'wasted' in training an Ising perceptron, compared to the case in which examples are random. Optimally stable spherical perceptrons may also be taught significantly more efficiently, and our results compare well with an existing numerical simulation.

1. Introduction

An ongoing problem in the field of neural networks is how to train a system to perform a task using correct examples (associations of 'questions' and 'answers'). The simplest algorithms select the questions at random [1–5] and use a variety of methods to design a perceptron from them, but this becomes increasingly inefficient as the number of examples is increased because most new questions give no new information. A more intelligent algorithm would *select* new questions on the basis of what had already been learnt so that the new information we would hope to obtain is maximized. One method of selection has been devised [6, 7], and applied to the simplest design of perceptron; the aim of this paper is to develop an analytical treatment of the learning process for more advanced perceptrons.

We are concerned with the problem of learning a linearly separable Boolean function T, which associates answer S_o with an N-vector question $\{S_i\}$: $i = 1, \ldots, N$, where $S_i \in \{-1, +1\}$. The function is defined by an N-vector $\boldsymbol{B} \in R^N$ so that

$$S_o = T(\boldsymbol{S}) = \operatorname{sign}(\boldsymbol{B} \cdot \boldsymbol{S}) \tag{1}$$

where we have defined the scalar product of two N-vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ as $\boldsymbol{x} \cdot \boldsymbol{y} = 1/N \sum_i x_i y_i$. Our perceptron will be given p question-and-answer examples $\{\boldsymbol{\xi}^\mu, \xi_o\}$, where $\xi_i^\mu \in \{-1, +1\}$, and $\mu = 1, \ldots, p$. The number of examples p will scale as αN and we will consider the limit $N \to \infty$ with α constant.

We will be using single-layer perceptrons defined by a single weight N-vector, $\boldsymbol{J}$, which gives an output determined by

$$S_1 = U(\boldsymbol{S}) = \operatorname{sign}(\boldsymbol{J} \cdot \boldsymbol{S}) . \tag{2}$$

At this point we can classify two sorts of perceptrons. 'Ising' perceptrons allow J_i to take only two values $\{+1, -1\}$ and are typically used [4] to solve problems in which

0305-4470/92/010113+09$03.50

B_i only takes the two values $\{+1, -1\}$. For problems in which B_i may take any real value, so that $\boldsymbol{B} \in R^N$, we should also allow J_i to take any real value [1–3] and this is called a 'spherical perceptron'. Our choice of $\boldsymbol{J}$, given p examples, we call $\boldsymbol{J}^{(p)}$. The simplest construction of a spherical perceptron is the Hebb rule

$$J_i^{(p)} = \frac{1}{\sqrt{N}} \sum_{\nu=1}^{p} \xi_o^\nu \xi_i^\nu . \tag{3}$$

An alternative approach is due to Gardner and Derrida [8], who search the whole N-dimensional J-space, looking for those choices of $\boldsymbol{J}^{(p)}$ which fulfil the equation

$$\kappa = \min_\mu \left\{ \left(\xi_o^\mu \sum_j J_j^{(p)} \xi_j^p \right) \left(\sum_j (J_j^{(p)})^2 \right)^{-1/2} \right\} \tag{4}$$

where the minimum is taken with respect to all the patterns $\{\boldsymbol{\xi}^\mu\}$. Any $\boldsymbol{J}^{(p)}$ for which $\kappa > 0$ will be able to correctly reproduce all the given answers ($U(\boldsymbol{\xi}^\mu) = T(\boldsymbol{\xi}^\mu)\ \forall \mu$); the $\boldsymbol{J}^{(p)}$ for which κ is greatest has 'optimal stability', and algorithms exist which will construct this $\boldsymbol{J}$, notably the MinOver algorithm [9].

One quantity of particular interest is the 'generalization probability', $G(\alpha)$, which is the probability that a random state $\boldsymbol{S}$ gives the correct output (i.e. $T(\boldsymbol{S}) = U(\boldsymbol{S})$). The generalization probability has been calculated [1] for rule (3) and has the following behaviour: $G(\alpha = 0) = 0.5$, which means that a random question has probability $\frac{1}{2}$ of being answered correctly; for higher α, G increases as $\boldsymbol{J}^{(p)}$ becomes closely aligned to $\boldsymbol{B}$. If the proportion of patterns presented becomes large, the system reaches perfect generalization, i.e. $G(\alpha \to \infty) = 1$. However, it has been found [3] that if more than $\alpha \approx 1$ patterns are presented the optimal stablity rule for the spherical perceptron gives a higher $G(\alpha)$ than the Hebb rule. In fact as $\alpha \to \infty$, $1 - G(\alpha) \sim \alpha^{-1/2}$ for the Hebb rule and $1 - G(\alpha) \sim \alpha^{-1}$ for the optimal perceptron.

Rather different behaviour has been observed for an 'Ising' perceptron [4, 5]. If we choose a value of κ and we use a standard technique (replica theory, [10]) to calculate the logarithm of the number of solutions to (4), the entropy of the answer, $S(\alpha, \kappa)$, we find that for higher values of α than a critical line $\alpha(\kappa)$ the entropy becomes negative, which is impossible since $\boldsymbol{J}^{(p)}$ can take only a finite number of values (2^N). For $\alpha \gtrsim \alpha_c(0)$, the only $\boldsymbol{J}^{(p)}$ which correctly stores all examples is $\boldsymbol{B}$, and the argument of [11] suggests that $G(\alpha)$ jumps suddenly to 1, implying perfect generalization. It was shown in [4] that $\alpha_c(0) = 1.245$; since at least N examples are needed to specify the N bits of $\boldsymbol{B}$, the minimum α at which the transition should occur is 1, and thus $0.245N$ examples (19.7%) are 'wasted'.

We gain some insight into the process of learning from a diagram of a two-dimensional section of the N-dimensional space containing $\boldsymbol{B}$ (figure 1). In the figure $\mathcal{A}$ is the $(N-1)$-dimensional hyperplane perpendicular to $\boldsymbol{B}$, and $\boldsymbol{\xi}^{\text{rand}}$ is a random example, which has the orthogonal hyperplane $\mathcal{Y}$. The output of $\boldsymbol{\xi}^{\text{rand}}$ with respect to rule T is $+1$, if $\boldsymbol{\xi}^{\text{rand}}$ and $\boldsymbol{B}$ lie on the same side of plane $\mathcal{A}$. Thus knowing the value of $T(\boldsymbol{\xi}^{\text{rand}})$ simply determines on which side of plane $\mathcal{Y}$ vector $\boldsymbol{B}$ lies. In order to constrain $\boldsymbol{B}$ into a very small subspace, we require a large number of examples lying close to plane $\mathcal{A}$. If we select examples randomly we must wait until sufficiently many with this property happen to occur.

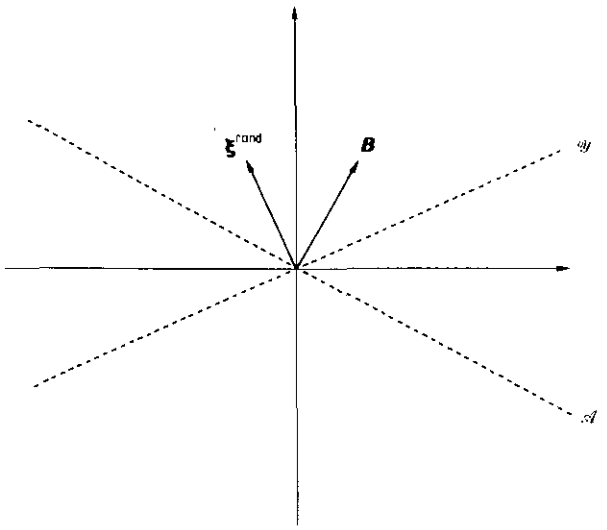

Figure 1. A section of the N-dimensional space containing $\boldsymbol{B}$, orthogonal to hyperplane $\mathcal{A}$. The vector $\boldsymbol{\xi}^{\text{rand}}$ is a random example in this space orthogonal to hyperplane $\mathcal{Y}$.

In [6] a method for speeding up the process was suggested. Figure 2 shows the same space, where we have marked $\boldsymbol{J}^{(p)}$, our best estimate for $\boldsymbol{B}$ after the presentation of p examples. We should therefore select our next question $\boldsymbol{\xi}^{p+1}$ to lie in the plane $\mathcal{Z}$ (perpendicular to $\boldsymbol{J}^{(p)}$), so that $\boldsymbol{\xi}^{p+1} \cdot \boldsymbol{B}$ is minimized. This implies that we choose $\boldsymbol{\xi}^{p+1}$ at random but with the constraint

$$\sqrt{N}\boldsymbol{\xi}^{p+1} \cdot \boldsymbol{J}^{(p)} = 0 \tag{5}$$

in the limit of $N \to \infty$. From [6] we know that the parameter x_p, defined by

$$x_p = \boldsymbol{\xi}^{p+1} \cdot \boldsymbol{B}\sqrt{N} \tag{6}$$

has a distribution

$$\rho(x_p) = \frac{1}{\sqrt{2\pi}\sin(\theta_p)} \exp\left(-\frac{x_p^2}{2\sin^2(\theta_p)}\right) \tag{7}$$

where we introduced θ_p as

$$\cos(\theta_p) = \boldsymbol{J}^{(p)} \cdot \boldsymbol{B} \equiv R(\alpha). \tag{8}$$

Equation (8) also defines the order parameter $R(\alpha)$ as the overlap between $\boldsymbol{B}$ and $\boldsymbol{J}$. We have used the normalization $\boldsymbol{J} \cdot \boldsymbol{J} = \boldsymbol{B} \cdot \boldsymbol{B} = 1$. From figure 1 we see that

$$G(\alpha) = 1 - \frac{\theta_p}{\pi} = 1 - \frac{1}{\pi}\cos^{-1}(R(\alpha)) \tag{9}$$

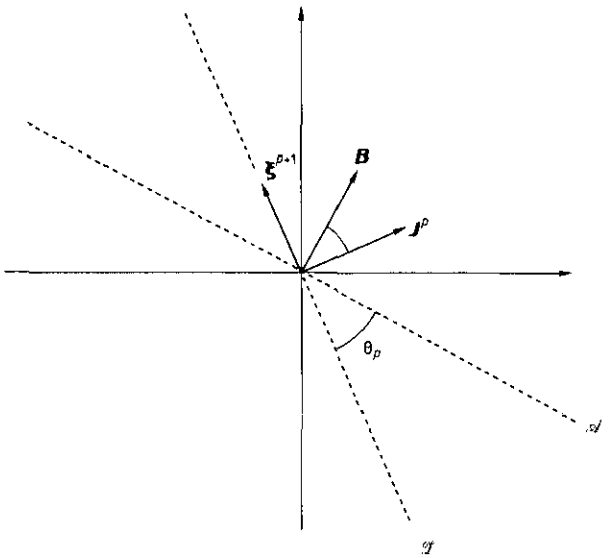

Figure 2. A section of the N-dimensional space containing $\boldsymbol{B}$, orthogonal to hyperplane $\mathcal{A}$. $\boldsymbol{J}^p$, orthogonal to hyperplane $\mathcal{Z}$, is our best estimate for $\boldsymbol{B}$ after presenting p examples. The angle between $\boldsymbol{J}^p$ and $\boldsymbol{B}$ is θ^p. ξ^{p+1} is chosen at random to lie in hyperplane $\mathcal{Z}$.

since it is the fraction of the N-dimensional space on the same side of planes $\mathcal{A}$ and $\mathcal{Z}$.

When the results of this algorithm were evaluated [6] for the 'spherical' perceptron using the Hebb rule, it was found that the resultant $G(\alpha)$ was higher than with any known setting of $\boldsymbol{J}^{(p)}$ using random examples. [6] also contains a numerical simulation for a spherical perceptron using the optimal stability rule which exhibits a further significant improvement. It is the purpose of this paper to develop a technique which produces an analytical expression for this result, and then to apply it to the Ising perceptron.

2. Spherical perceptrons

As usual we begin by considering the volume of $\{\boldsymbol{J}\}$-space satisfying (4) for any given κ, which is

$$Z = \int \left(\prod_i \mathrm{d}J_i \right) \prod_{\mu=1}^{p} \Theta \left(\frac{\xi_o^\mu}{\sqrt{N}} \sum_j J_j^{(p)} \xi_j^\mu - \kappa \right) \delta \left(\sum_j J_j^2 - N \right). \quad (10)$$

and which shrinks to zero for any α as $\kappa \to \kappa(\alpha)$, the maximal stability. The order parameters $\kappa(\alpha)$ and $R(\alpha)$ of the system will be found, as in [3], from $\langle \ln Z \rangle$, where the average is taken over all the patterns which might be generated during the process. We perform the average using the replica method [8] and working with Z^n.

Although we expect the angle θ_p between $\boldsymbol{J}^{(p)}$ and $\boldsymbol{B}$ to decrease as p increases, the hyperplane perpendicular to $\boldsymbol{B}$ is an $(N-1)$-dimensional subspace in which we expect

the component of $\boldsymbol{J}$ perpendicular to $\boldsymbol{B}$ to vary widely. Similarly the hyperplane perpendicular to $\boldsymbol{J}^{(p)}$, from which $\boldsymbol{\xi}^{p+1}$ is randomly chosen, is $(N-1)$-dimensional. Hence we make the ansatz that only correlations between the $\{\boldsymbol{\xi}^{\mu}\}$ in the direction of $\boldsymbol{B}$ are of significance, and therefore that any $\boldsymbol{\xi}^{\mu}$ satisfying (6) and (7) is equally likely to be realized. *Thus we assume that the average over $\boldsymbol{\xi}^{\mu+1}$ with constraint (5) is equivalent to the average with constraints (6) and (7)*, despite the fact that (6) and (7) do not themselves imply (5). This assumption was implicit in [6].

For a given x^{μ}, the number of patterns $\{\boldsymbol{\xi}^{\mu+1}\}$ satisfying (5) is

$$\mathrm{Tr}_{\{\boldsymbol{\xi}^{\mu+1}\}}\delta\left(\frac{1}{\sqrt{N}}\sum_i B_i\xi_i^{\mu+1}-x_{\mu}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_{\mu}^2}{2}+N\ln 2\right)\equiv m(x_{\mu}). \tag{11}$$

Thus the average of any function $\hat{A}$ over all the $\{\boldsymbol{\xi}^{\mu+1}\}$ satisfying (6) for a given x_{μ} is

$$\frac{1}{m(x_{\mu})}\mathrm{Tr}_{\{\boldsymbol{\xi}^{\mu+1}\}}\hat{A}\delta\left(x_{\mu}-\frac{1}{\sqrt{N}}\sum_i B_i\xi_i^{\mu+1}\right). \tag{12}$$

But equation (7) gives us the probability distribution for x_{μ} for every θ_{μ}. It follows that the operator which, working on $\ln Z$, produces the average over the $(\mu+1)$th pattern is

$$\frac{1}{\sin\theta_{\mu}}\int \mathrm{d}x\exp\left[\frac{x^2}{2}\left(1-\frac{1}{\sin^2\theta_{\mu}}\right)-N\ln 2\right]\mathrm{Tr}_{\{\boldsymbol{\xi}^{\mu+1}\}}\delta\left(x-\frac{1}{\sqrt{N}}\sum_i B_i\xi_i^{\mu+1}\right). \tag{13}$$

Applying this result and performing the trace in the manner of [3] we obtain, using the notation of [3],

$$\begin{aligned}\langle Z^n\rangle=&\int\left(\prod_{a<b}\mathrm{d}q_{ab}\right)\left(\prod_a \mathrm{d}R_a\right)\left(\prod_{ja}\mathrm{d}J_{ja}\right)\exp\left(\sum_{\mu=1}^{p}\Phi_{\mu}(q_{ab},R_a)\right)\\&\times\prod_{a<b}\delta\left(\sum_j J_{ja}J_{jb}-Nq_{ab}\right)\prod_a\delta\left(\sum_j J_{ja}B_j-NR_a\right)\delta\left(\sum_j J_{ja}^2-N\right)\end{aligned} \tag{14}$$

where we have introduced

$$\begin{aligned}\exp(\Phi_{\mu})\equiv&\int_{\kappa}^{\infty}\left(\prod_a\frac{\mathrm{d}\lambda_a}{2\pi}\right)\int\left(\prod_a\mathrm{d}x_a\right)\int\frac{\mathrm{d}u}{\sqrt{2\pi}\sin\theta_{\mu}}\exp\left(-\frac{u^2}{2\sin^2\theta_{\mu}}\right)\\&\times\exp\left(-\sum_{a<b}x_a x_b(q_{ab}-R_aR_b)-\tfrac{1}{2}\sum_a x_a^2(1-R_a^2)\right.\\&\left.+\mathrm{i}\sum_a x_a(R_a|u|-\lambda_a)\right).\end{aligned} \tag{15}$$

We can now replace

$$\sum_{\mu=1}^{p}\Phi_{\mu}(q_{ab},R_a)\to N\int_0^{\alpha}\mathrm{d}\alpha'\phi(q_{ab},R_a,\alpha') \tag{16}$$

where $\phi(q_{ab}, R_a, \alpha') \equiv \Phi_\mu(q_{ab}, R_a)$ when $\alpha' = \mu/N$. The error in making this replacement is not extensive and may thus be ignored in the $N \to \infty$ limit.

The rest of the calculation proceeds in the manner of [3]: we assume replica symmetry; perform the integrals over conjugate momenta using the saddle point method; take the $n \to 0$ limit; and finally allow the volume in the J-space to shrink to zero by letting $q \to 1$. We obtain as our final result the self-consistent equations for $\kappa(\alpha)$ and $R(\alpha)$, in terms of the $R(\alpha')$ for $\alpha' < \alpha$:

$$\int_0^\alpha \mathrm{d}\alpha' \int_{\mathcal{T}} \mathrm{d}u\mathrm{D}z \frac{\exp\{-u^2/[2(1-R^2(\alpha'))]\}}{[2\pi(1-R^2(\alpha'))]^{1/2}} \Upsilon^2(R(\alpha), \kappa(\alpha), u, z) = 1 - R^2(\alpha) \tag{17}$$

$$\frac{\partial}{\partial R(\alpha)} \left(\int_0^\alpha \mathrm{d}\alpha' \int_{\mathcal{T}} \mathrm{d}u\mathrm{D}z \frac{\exp\{-u^2/[2(1-R^2(\alpha'))]\}}{[2\pi(1-R^2(\alpha'))]^{1/2}} \Upsilon^2(R(\alpha), \kappa(\alpha), u, z) \right) = -2R(\alpha) \tag{18}$$

where, as a shorthand, we have introduced the function

$$\Upsilon(R(\alpha), \kappa(\alpha), u, z) \equiv \kappa(\alpha) - z\sqrt{1-R^2(\alpha)} - R(\alpha)|u|. \tag{19}$$

The region $\mathcal{T}$ of integration is given by $\Upsilon(R(\alpha), \kappa(\alpha), u, z) > 0$.

The equations are very similar to those of [3], but the difference has an appealing physical interpretation. If we interpret u as a noise we find that it is Gaussian for every value of α' but decreases in width as $R(\alpha')$ increases (instead of remaining a constant width as in [3]). This corresponds to $\{\boldsymbol{\xi}^\mu\}$ placing on average a more stringent constraint per example on $\boldsymbol{B}$ as μ increases. Numerical solution of equations (17) and (18) is somewhat difficult because to avoid integrating around $\alpha \sim 0$, where κ diverges, we must estimate how R behaves in this region. Our solutions are plotted in figure 3 as line 1 with the vertical lines to show our estimate of precision, ± 0.01, and with the numerical results [6] as dots for comparison; the agreement between results and analysis is excellent, well within the experimental error. The prediction using random examples is shown as line 2. A usual question when applying the replica method is whether the replica symmetric assumption is valid [10]. In this case we believe that it is, just as in the original spherical model [8], and this belief is supported by the quality of agreement with the numerical results.

3. Ising perceptrons

Applying the same argument to the problem of Ising perceptrons, we have obtained an expression for the entropy

$$S = \langle \ln Z \rangle_\xi = \mathrm{Extr}_{R,f,g,q} \left[-gR - \frac{f}{2}(1-q) + \int_{-\infty}^{\infty} \mathrm{D}t \ \ln 2\cosh(t\sqrt{f} + g) \right.$$
$$\left. + 2\int_0^\alpha \mathrm{d}\alpha' \int_{-\infty}^{\infty} \mathrm{D}t \, H\left(\frac{tR\gamma}{\sqrt{q-R^2}} \right) \ln H\left(\frac{\kappa + t\sqrt{q - R^2(1-\gamma^2)}}{\sqrt{1-q}} \right) \right] \tag{20}$$

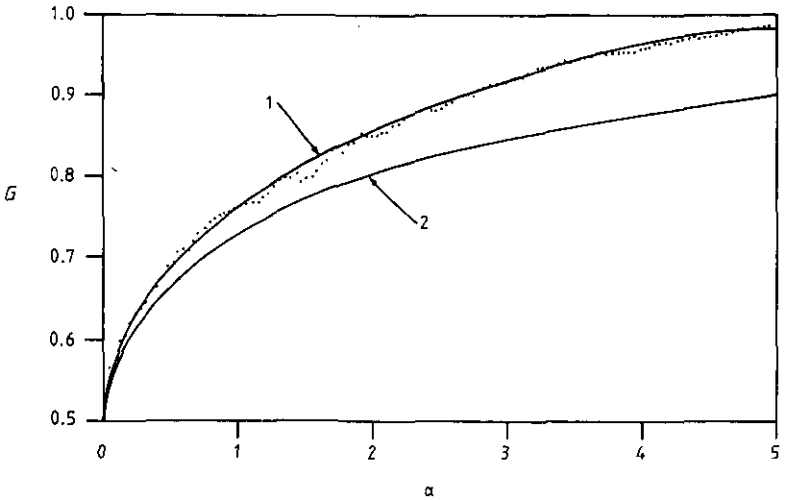

Figure 3. The generalization ability $G(\alpha)$ against α. Line 1 is the prediction of this paper. The dots show the results of a numerical simulation with $N = 50$ [6]. Line 2 is the prediction using a random selection of examples [3].

where $\gamma = 1$ for unselected examples, and $\gamma = \sqrt{1 - R^2(\alpha')}$ for selected ones. This form was derived in [4] with $\gamma = 1$ and $\kappa = 0$, so that any $\boldsymbol{J}^{(p)}$ was chosen which correctly stored the p examples. We assume that, as in [4], and as explained above in the introduction, there is a phase change to perfect generalization when S becomes zero.

Figure 4 shows the consequences of this equation, and the inset is an expanded version of the region of interest. Line 1 is the one derived in [4] with unselected examples using the ansätze $q = R$ and $f = g$ (we have checked the validity of these ansätze by a search of the whole space); it shows a first-order transition to perfect generalization ($G(\alpha) = R = 1$) at $\alpha \approx 1.245$ from $R \approx 0.697$ and $G(\alpha) \approx 0.756$. Line 2 shows the optimal stability result, where at each α the value of κ is increased until $S = 0$; thus the two lines coincide at their endpoints. In general points on line 2 have $q > R$ and $f > g$, since enforcing optimal stability reduced the available volume in J-space, but not necessarily around $\boldsymbol{B}$. It is worth pointing out that line 2 marks the overlap just *before* the freezing transition which occurs as κ is increased (when the available volume in J-space contains just one point). To find out the overlap with $\boldsymbol{B}$ of the state it freezes into (a line higher than 2) we would have to use a more subtle analysis, presumably with first step replica symmetry breaking; we know only that at the end of line 2 the jump must be to perfect generalization.

Line 3 shows the results if examples are selected but optimal stability is not enforced, $\kappa = 0$. The phase transition to perfect generalization now occurs at $\alpha \approx 1.173$, $R \approx 0.661$ and $G(\alpha) \approx 0.730$, a 30% advance, compared to unselected examples, towards the theoretical minimum bound of $\alpha = 1$.

Our best result, line 4, is obtained by at each step choosing the optimally stable $\boldsymbol{J}^{(p)}$ and selecting the next example. On this occasion the transition to perfect generalization occurs at $\alpha \approx 1.145$ and $G(\alpha) \approx 0.727$, which is 41% closer to the minimum bound than learning using random examples. This represents an 8% reduction of the total number of examples required by what is already an efficient training algorithm.

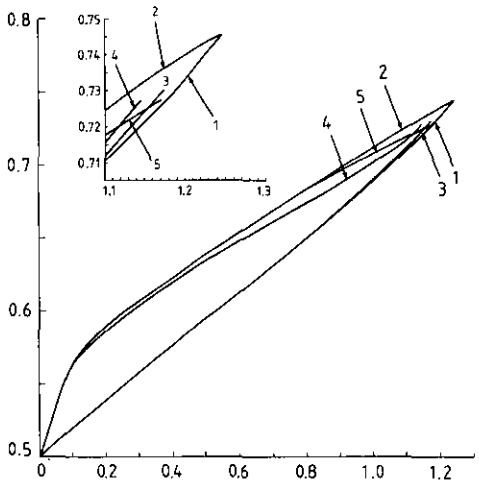

Figure 4. Selecting examples for an Ising perceptron. Line 1, unselected examples, no optimization; line 2, unselected examples, optimal stability; line 3, selected examples but no optimization; line 4, selected examples and optimal stability; line 5, optimal stability with examples selected for $\alpha > 0.7$. All lines show discontinuous transitions to $G = 1.0$ for $1.1 < \alpha < 1.3$. This interesting region is shown expanded as an inset.

Note that it is the $R(\alpha)$ in line 4 which is used iteratively to define γ in (20); this corresponds to selecting examples at each value of α using a $\boldsymbol{J}$ just before the freezing transition. The RSB analysis, however, would straightforwardly allow us to analyse the case of selecting examples after freezing.

Line 4 lies below line 2 for $\alpha > 0$. This, we suggest, is because line 4 marks the freezing transition between the available volume containing two points and containing one. Selecting examples would mean that the volume in which $\boldsymbol{B}$ is confined is bordered by fewer planes and so is less convex—in the sense that, for the same volume, the average distance between two points is greater. Thus the average overlap with $\boldsymbol{B}$ is lower before freezing. As explained above, line 4 should not be taken as an indication of the position of the line (not shown) marking the overlap of the state into which $\boldsymbol{J}$ freezes; we know only that this line touches $G = 1$ at $\alpha \approx 1.145$ and thus, at least here, is above the line marking the overlap of the frozen state using unselected examples.

It is interesting to observe that if we begin to select examples some way into the learning process then the improvement in efficiency is only very slightly reduced. Line 5 shows the result of always enforcing maximum stability, but only selecting examples for $\alpha > 0.7$. The first-order transition now occurs at $\alpha \approx 1.170$, which is better than the result of always selecting examples but never enforcing maximum stability (line 3).

We expect that our results are stable with respect to the breaking of replica symmetry, since we are working in the region of α rather lower than the dA–T line [4].

4. Conclusion

We have shown that by accepting a very plausible ansatz the problem of learning with continually selected examples can be treated, and our technique has been amply justified by agreement with an existing numerical simulation for the spherical perceptron. In Ising perceptrons we have predicted an improvement in efficiency over the previous algorithm which takes us 41% nearer to the theoretical maximum bound.

The technique of selecting examples has many natural further applications, of which the most obvious is to perceptrons with more sophisticated geometries than our simple one-layer machine. Networks with hidden layers are capable of solving problems in which the Boolean function T is not linearly separable (for example, the parity problem), and these perceptrons have recently been studied using random examples [12].

A further generalization would be to quite different sorts of learning problems. The perceptrons of [13], for example, learn to classify inputs according to their Hamming distance from a set of prototype patterns; selection of examples may well lead again to significant advances in efficiency.

Acknowledgments

We thank the SERC for financial support. One of us (AR) would like to thank the *Studienstiftung des deutschen Volkes* for the award of a special one year travelling scholarship. We thank Professor David Sherrington for carefully reading the final version of the manuscript.

References

[1] Vallet F 1989 *Europhys. Lett.* **8** 747
[2] Vallet F, Cailton J and Refregier P 1989 *Europhys. Lett.* **9** 315
Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific)
[3] Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581
[4] Györgyi G 1990 *Phys. Rev.* A **41** 7097
[5] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683
[6] Kinzel W and Ruján P 1990 *Europhys. Lett.* **13** 473
[7] Baum E B 1991 *IEEE Trans. Neural Networks* **2** 5
[8] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983
[9] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[10] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[11] Krauth W and Mézard M 1989 *J.Physique* **50** 3067
[12] Barkai E and Kanter I 1991 *Europhys. Lett.* **14** 107
[13] Hansel D and Sompolinsky H 1990 *Europhys. Lett.* **11** 687