On-line learning in a discrete state space

1998 J. Phys. A: Math. Gen. 31 L27

(http://iopscience.iop.org/0305-4470/31/1/004)

J. Phys. A: Math. Gen. **31** (1998) L27–L30. Printed in the UK PII: S0305-4470(98)84472-7

LETTER TO THE EDITOR

On-line learning in a discrete state space

W Kinzel and R Urbanczik

Institut für theoretische Physik, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany

Received 22 May 1997, in final form 6 October 1997

Abstract. On-line learning of a rule given by an N-dimensional Ising perceptron is considered for the case when the student is constrained to take values in a discrete state space of size L^N. For $L = 2$ no on-line algorithm can achieve a finite overlap with the teacher in the thermodynamic limit. However, if L is on the order of $\sqrt{N}$, Hebbian learning does achieve a finite overlap.

Artificial neural networks are usually trained by a set of examples [1]. After the training phase such a network ('student') has achieved some knowledge about the rule ('teacher') which has generated the examples. The difference between the outputs of the student and the teacher for a random input vector defines the generalization error.

There are two basic kinds of training algorithms: (1) in the batch mode the complete set of examples is stored and iteratively used to change the synaptic weights of the student network; (2) in the on-line mode each example is used only once, at each training step a new example is presented and the synaptic weights are changed according to some algorithm.

The analysis of on-line algorithms using methods of statistical mechanics [2–6] has shown that this is a powerful and versatile approach to learning problems. To our knowledge, however, only continuous couplings have so far been considered. However, for hardware implementations it would be extremely useful to design algorithms which work in a discrete space of synaptic weights. It is not known whether on-line algorithms work at all for weights which have a limited number L of possible values. Here we show for a simple case that generalization is only possible if L is of the order of $\sqrt{N}$, where N is the size of the network. Hence for a fixed depth L of the synaptic weights, on-line learning will not generalize at all in the thermodynamic limit. This is in contrast to batch learning, where for $L = 2$ a transition to perfect generalization is found at a critical size of the training set [7, 8].

We consider the perhaps simplest learning scenario in which the teacher is a perceptron with N binary couplings $B_i \in \{-1, 1\}$. In on-line learning, the student perceptron with weight vector J receives at each time step an N-dimensional input ξ and the classification bit $\sigma_B(\xi) \in \{-1, 1\}$ provided by the teacher B. The task is to find a mapping, $J' = f(J, \xi, \sigma_B(\xi))$ which updates the student J, our current approximation of B, based on this information. Of course, J' should be an improved approximation. Under very general conditions, we show in this letter that no such mapping exists if J and J' are confined to lie, as the teacher is, in the set $\{-1, 1\}^N$. In a second step, we consider Hebbian learning in a discretized state space of size L^N, and determine the generalization behaviour as a function of $\lambda = L/\sqrt{N}$.

The classification of ξ is given by $\sigma_B(\xi) = \text{sign}(B^T\xi)$. Hence the quality of the approximation provided by a student J can be defined via the overlap $R = N^{-1}B^T J$ with the teacher. Since the students have binary components, it is convenient to have the update

0305-4470/98/010027+04$19.50

rule f specify at which sites the sign should be flipped to obtain the updated weight vector J'. So $J_i' = J_i f_i(J, \xi, \sigma_B(\xi))$ and the f_i take values in $\{-1, 1\}$. The update rule will be useful if it improves on our current state, that is if

$$B^T J' = \sum_{i=1}^{N} B_i J_i f_i(J, \xi, \sigma_B(\xi)) > B^T J. \tag{1}$$

Of course, f cannot have any built-in knowledge about the teacher but must infer information about B from the current pattern. Formally, this can be enforced by requiring that f be useful not just for the single teacher B but, on average, for teachers which have the same overlap as B with J. Denoting by $\langle \ldots \rangle_{B|B^T J = NR}$ the average over the uniform distribution on the set of teachers which have overlap R with J, a useful f must thus fulfill

$$\left\langle \sum_{i=1}^{N} B_i J_i f_i(J, \xi, \sigma_B(\xi)) \right\rangle_{B|B^T J = NR} > NR. \tag{2}$$

By a gauge transformation, the left-hand side may be written as

$$\left\langle \sum_{i=1}^{N} B_i f_i(J, \xi, \sigma_B(\xi^*)) \right\rangle_{B|\sum_i B_i = NR}$$

where ξ^* is given by $\xi_i^* = J_i \xi_i$. Using the fact that for the Heaviside step function θ, $1 = \theta(\sigma_B(\xi^*)) + \theta(-\sigma_B(\xi^*))$, we may rewrite (2) as

$$\sum_{\sigma \in \{-1,1\}} \sum_{i=1}^{N} f_i(J, \xi, \sigma) \langle B_i \theta(\sigma B^T \xi^*) \rangle_{B|\sum_i B_i = NR} > NR. \tag{3}$$

Under mild conditions on ξ, one finds that

$$\langle B_i \theta(\sigma B^T \xi^*) \rangle_{B|\sum_i B_i = NR} \geqslant 0 \tag{4}$$

for any positive R in the limit of large N. Consequently, the left-hand side of (3) is maximized by choosing $f_i(J, \xi, \sigma) = 1$, and the best we can do is to keep the weight vector J fixed.

There are some special cases where (4) is not true. If just a single component of ξ is non-zero, then $\sigma_B(\xi)$ will of course give us the corresponding component of B and one can achieve $R = 1$ by asking N such questions. However, it is hard so see how such a strategy might be extended to the case of a noisy teacher.

For more generic patterns, however, the ξ_i will be of similar magnitude. Furthermore, ξ will only have a small overlap with J, that is $m = \sum_i \xi_i J_i / |\xi|$ will be of order 1. Then for large N, and consequently small $\xi_i / |\xi|$, the left-hand side of (4) may be evaluated using the central limit theorem and yields

$$\langle B_i \theta(\sigma B^T \xi^*) \rangle_{B|\sum_i B_i = NR} = RH\left(-\sigma m \frac{R}{\sqrt{1-R^2}} - \frac{\sqrt{1-R^2}}{R} \frac{\sigma \xi_i J_i}{|\xi|} \right) \tag{5}$$

which is positive. So if the components of ξ are picked independently from distributions having bounded ratios of their variances, the fraction of inputs for which (4) is violated decreases exponentially with N.

An even stronger statement can be made for binary inputs, $\xi_i \in \{-1, 1\}$. Then the large N expansion yielding (5) can only be wrong, if the input is correlated with the student ($|m| \gg 1$). However, for this case (4) may be verified by evaluating its left-hand side with the saddle-point method. Consequently, for binary inputs, on-line learning is impossible even if queries [5] are allowed.

As it is possible to learn on-line with continuous couplings, the question arises what the numerical depth of the couplings must be for on-line learning to succeed. We thus consider a situation where the J_i are constrained to lie in the set $\{1, 2, \ldots, L\}$, still with a binary teacher. A weight vector J is then taken to represent an estimate $\tilde{B}$ of B via $\tilde{B}_i = \mathrm{sign}(J_i - L/2)$. For randomly chosen binary inputs, Hebbian learning may be applied to J by truncating to the allowed range of values:

$$J_i' = \begin{cases} J_i + \xi_i \sigma_B(\xi) & \text{if } J_i + \xi_i \sigma_B(\xi) \in \{1, \ldots, L\} \\ J_i & \text{else.} \end{cases} \tag{6}$$

The increments $\xi_i \sigma_B(\xi)$ are not independent over the sites i but their covariances do decay as $1/N$. So for large N the sites will approximately decouple, and we are left with a biased random walk on each site. The bias is given by

$$\langle \xi_i \sigma_B(\xi) \rangle = B_i \sqrt{\frac{2}{\pi N}} \tag{7}$$

where $\langle \ldots \rangle$ is an average over random vectors ξ.

Let $p_l(t)$ denote the probability that $J_1 = l$ after t iterations of (6) and assume that $B_1 = 1$. Then

$$\begin{aligned} p_1(t+1) &= r p_1(t) + r p_2(t) \\ p_l(t+1) &= g p_{l-1}(t) + r p_{l+1}(t) \qquad l = 2, \ldots, L-1 \\ p_L(t+1) &= g p_{L-1}(t) + g p_L(t) \end{aligned} \tag{8}$$

where $r + g = 1$ and $g = 1/2 + 1/\sqrt{2\pi N}$ for large N. The stationary solution p^{s} of (8) is $p_l^{\mathrm{s}} \propto (g/r)^l$. Thus for large N the asymptotic overlap R^{s} between the estimate $\tilde{B}$ and the teacher will approach zero if L is fixed. For $L = \lambda\sqrt{N}$, however, one finds

$$R^{\mathrm{s}} = 1 - \frac{2}{1 + \mathrm{e}^{\sqrt{8/\pi}\lambda}}. \tag{9}$$

The time needed to approach the stationary distribution will scale linearly with N for fixed λ. So let $R(\alpha)$ be the overlap after αN steps, assuming that initially $J_i = L/2$. The time evolution of R may then be calculated using the explicit formulae for the powers of the transition matrix of the random walk (8) given in [9]. One finds

$$R(\alpha) = R^{\mathrm{s}} - 4\sqrt{2/\pi}\,\mathrm{e}^{-\alpha/\pi} \sum_{k=0}^{\infty} \mathrm{e}^{-(\pi^2\alpha/2\lambda^2)(2k+1)^2} \frac{\lambda}{(2/\pi)\lambda^2 + \pi^2(2k+1)^2}. \tag{10}$$

The resulting dependence of the overlap on λ (for fixed α) is non-monotonic as shown in figure 1. For large α the sum in the above expression is dominated by the first term and the difference between R and its asymptotic value R^{s} decays exponentially. The relaxation time is given by

$$\tau = \frac{2\lambda^2\pi}{2\lambda^2 + \pi^3} \tag{11}$$

and increases with λ. Thus for small values of λ the overlap will increase from zero up to R^{s} quickly but, on the other hand, this asymptotic value will not be close to unity because of (9). This competition leads to the non-monotonic behaviour found in figure 1. The relaxation time τ stays finite for large λ, and thus in the limit of a large number of examples the optimal choice is $\lambda \to \infty$.

To find the behaviour for $L \gg \sqrt{N}$, we need to take the limit $\lambda \to \infty$ in (10), that is, replace the sum over k by an integral. This yields

$$R(\alpha) = 1 - 2H(\sqrt{2\alpha/\pi}) \tag{12}$$

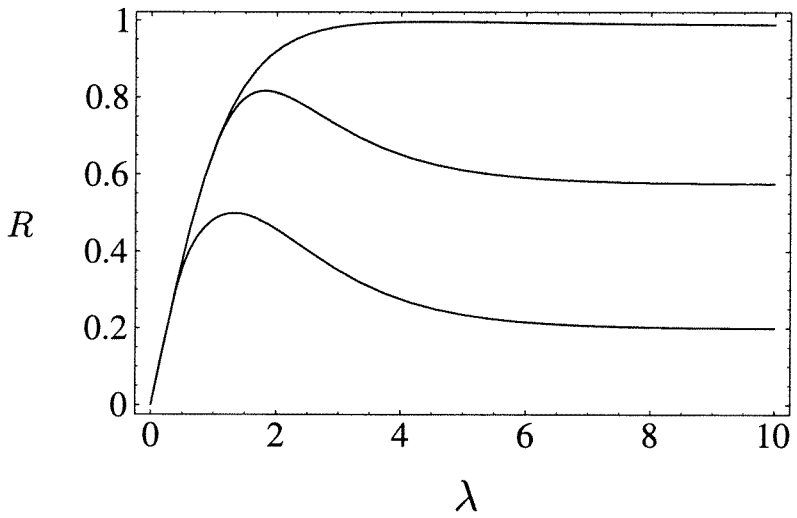

Figure 1. Overlap R achieved by Hebbian learning using L different weight values per coupling, $L = \lambda\sqrt{N}$. The curves are, from top to bottom, for $\alpha = 10$, $\alpha = 1$ and $\alpha = 0.1$

the result found in [10] for the case where one applies Hebb's rule to continuous couplings and clips in the end.

For $L > 2$, here we have considered only simple Hebbian learning. However, since αN examples will be needed to achieve good generalization, we believe that one cannot improve on the scaling, $L = \lambda\sqrt{N}$, by using a different algorithm.

One of the authors (WK) would like to thank Ido Kanter for useful discussions. The work of RU was supported by the Deutsche Forschungsgemeinschaft (DFG).

References

[1] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)
[2] Biehl M and Schwarze H 1993 Learning drifting concepts with neural networks *J. Phys. A: Math. Gen.* **26** 2651
[3] Kim J and Sompolinsky H 1996 On-line Gibbs learning *Phys. Rev. Lett.* **76** 3021–4
[4] Kinouchi O and Caticha N 1992 Optimal generalization in perceptrons *J. Phys. A: Math. Gen.* **25** 6243
[5] Kinzel W and Rujan P 1990 Improving a networks generalization ability by selecting examples *Europhys. Lett.* **13** 473–8
[6] Saad D and Solla S 1995 Exact solution for on-line learning in multilayer neural networks *Phys. Rev. Lett.* **74** 4337–40
[7] Sompolinsky H, Tishby N and Seung H S 1990 Learning from examples in large neural networks *Phys. Rev. Lett.* **65** 1683–6
[8] Györgyi G 1990 First-order transition to perfect generalization in a neural network with binary synapses *Phys. Rev.* A **41** 7097–100
[9] Feller W 1951 *Probability Theory and its Applications* vol I (New York: Wiley)
[10] Van den Broeck C and Bouten M 1993 Clipped-Hebbian training of the perceptron *Europhys. Lett.* **22** 223–9