Perceptron-like Large Margin Classifiers

Petroula Tsampouka and John Shawe-Taylor

School of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK
e-mail: {pt04r , jst}@ecs.soton.ac.uk

ABSTRACT

We consider perceptron-like algorithms with margin in which the standard classification condition is modified to require a specific value of the margin in the augmented space. The new algorithms are shown to converge in a finite number of steps and used to approximately locate the optimal weight vector in the augmented space following a procedure analogous to Bolzano's bisection method. We demonstrate that as the data are embedded in the augmented space at a larger distance from the origin the maximum margin in that space approaches the maximum geometric one in the original space. Thus, our algorithmic procedure could be regarded as an approximate maximal margin classifier. An important property of our method is that the computational cost for its implementation scales only linearly with the number of training patterns.

1 Introduction

Rosenblatt's perceptron [7] is the simplest on-line learning algorithm for binary linear classification [4]. In its original form it does not insist on finding a non-zero margin of the dataset from the solution hyperplane but cares only for correct classification. It is generally believed that the larger the margin of the dataset the greater is the generalisation ability of the learning machine. For this reason a variant of the perceptron algorithm, known as the perceptron with margin, was introduced. This new algorithm converges to a solution possessing a non-zero margin which, however, is an unknown fraction of the maximum existing one. The problem of finding the optimal margin hyperplane has been successfully addressed only with the advent of the simplest Support Vector Machine (SVM) [8,3], the maximal margin classifier [2].

Our purpose in the present work is to address the problem of maximal margin classification using the less time consuming, compared to SVMs, perceptron-like algorithms. We work in an augmented by one additional dimension space [4] in which we embed the data by placing them at a distance ρ in the extra dimension and replace the classification condition of the perceptron with a new one which insists on a specific value of the margin in this augmented space. We show that the algorithms with the modified condition converge in a finite number of steps and use them to approximately locate the solution with maximum margin in the augmented space. Our search is performed employing a procedure which resembles Bolzano's bisection method. Finally, we derive an upper bound on the geometric margin involving the maximum margin in the augmented space and the displacement distance ρ of the data in the additional dimension. From this upper bound follows that in the limit $\rho \to \infty$ the maximum margin in the augmented space approaches the maximum geometric one in the original space. Thus, our algorithmic procedure could be considered an approximate maximal margin classifier.

In the process of proving convergence of the algorithms with the new type of condition we found useful to introduce the notion of stepwise convergence, the property of the algorithms to approach in each step the optimal solution vector. This led to a unified approach in establishing convergence of a large class of algorithms with additive perceptron-like update rules irrespectively of the type of the classification condition.

The organisation of the paper is as follows. Section 2 contains our theoretical analysis and consists of 3 subsections. The first subsection deals with the convergence of algorithms with standard margin condition while the second subsection is concerned with the convergence of the new algorithms with fixed margin condition which fall into two categories depending on whether the length of the weight vector is free or fixed. The last subsection

contains our considerations which lead to an estimate of the geometric margin. In section 3 we describe the algorithmic implementation aiming at an approximate determination of the maximum margin. Finally, section 4 contains our conclusions.

2 Theoretical analysis

In what follows we make the assumption that we are given a training sample which, if not initially linearly separable, by an appropriate feature mapping into a space of a higher dimension [1,2] can be classified into two categories by a linear classifier. This higher dimensional space in which the patterns are linearly separable will be our original space. By adding to the original space one additional dimension and placing all patterns in the same position in that dimension we construct an embedding of our data in the so-called augmented space.

In this paper we study algorithms that update the augmented weight vector $\bar{a}_t$ by adding a suitable positive amount in the direction of the misclassified (according to an appropriate condition) training pattern $\bar{y}_k$. In the general case this amount exhibits a dependency on the current step which could be due to the current weight vector and/or the misclassified training pattern which is presented to the algorithm at the specific step. As such, this amount should be considered a function of time and be denoted by f_t. For the special case of the perceptron algorithm $f_t = 1$. Thus, the general form of the update rule is

$$\bar{a}_{t+1} = \bar{a}_t + \eta f_t \bar{y}_k, \tag{2.1}$$

where η is the learning rate and should be considered a constant parameter of the algorithm. Each time the predefined condition is satisfied by a training pattern the algorithm proceeds to the update of the weight vector. Throughout our discussion a reflection with respect to the origin in the augmented space of the negative label patterns is understood in order to allow for a common classification condition for both categories of patterns [4].

2.1 Algorithms with the standard margin condition

First we examine algorithms in which the misclassification condition that should be checked takes the form

$$\bar{y}_k \cdot \bar{a}_t \leq b, \tag{2.2}$$

where b is a positive parameter. The condition characterising optimally correct classification of the training patterns by a weight vector $\bar{u}$ of unit norm in the augmented space is

$$\bar{y}_k \cdot \bar{u} \geq \gamma_d \quad \forall k. \tag{2.3}$$

The quantity γ_d, which we call the optimal directional margin, is defined by the relation

$$\gamma_d = \max_{\bar{u}} \min_k \{\bar{y}_k \cdot \bar{u}\}. \tag{2.4}$$

From its definition it becomes obvious that γ_d is bounded from above by $r = \min_k \|\bar{y}_k\|$. The optimal directional margin determines the maximum distance from the origin in the augmented space of the hyperplane normal to $\bar{u}$ placing all training patterns on the positive side. In the determination of this hyperplane only the direction of $\bar{u}$ is exploited with no reference to its projection onto the original space. As a consequence the above maximum margin in the augmented space is not necessarily realised with the same weight vector that gives rise to the optimal geometric margin in the original space.

We analyse the algorithms with the general update rule (2.1) by calculating an upper bound on the number of updates until the solution is found. To achieve this we resort to an extension of Novikoff's theorem [6] for which it is required that f_t be positive and bounded, i.e.

$$0 < f_{min} \leq f_t \leq f_{max}. \tag{2.5}$$

Throughout we use the shorthand notation $R = \max_k \|\bar{y}_k\|$. From the difference of the inner products of $\bar{u}$ with the weight vector $\bar{a}_t$ at successive time steps we have

$$\bar{a}_{t+1} \cdot \bar{u} - \bar{a}_t \cdot \bar{u} = \eta f_t \bar{y}_k \cdot \bar{u} \geq \eta f_{min} \gamma_d. \tag{2.6}$$

A repeated application of Eq. (2.6) with the assumption that $\bar{a}_t$ is initially set to zero implies that

$$\|\bar{a}_t\| \geq \bar{a}_t \cdot \bar{u} \geq \eta f_{min} \gamma_d t, \tag{2.7}$$

which gives us a lower bound on $\|\bar{a}_t\|$. By calculating the difference of the squared norms of the weight vectors in consecutive steps we obtain

$$\|\bar{a}_{t+1}\|^2 - \|\bar{a}_t\|^2 = \eta^2 f_t^2 \|\bar{y}_k\|^2 + 2\eta f_t \bar{y}_k \cdot \bar{a}_t \leq \eta^2 f_{max}^2 R^2 + 2\eta f_{max} b. \tag{2.8}$$

A repeated application of Eq. (2.8) leads to the following upper bound on $\|\bar{a}_t\|$

$$\|\bar{a}_t\| \leq \sqrt{(\eta^2 f_{max}^2 R^2 + 2\eta f_{max} b) t}. \tag{2.9}$$

Combining Eqs. (2.7) and (2.9) we get the squeezing relationship

$$\eta f_{min} \gamma_d t \leq \bar{a}_t \cdot \bar{u} \leq \|\bar{a}_t\| \leq \sqrt{(\eta^2 f_{max}^2 R^2 + 2\eta f_{max} b) t} \tag{2.10}$$

from which the following time bound for convergence is derived

$$t \leq t_N \equiv 2\frac{f_{max}}{f_{min}}\left(\frac{1}{2}\frac{f_{max}}{f_{min}}\frac{R^2}{\gamma_d^2} + \frac{1}{\eta f_{min}}\frac{b}{\gamma_d^2}\right). \tag{2.11}$$

A very desirable property of an algorithm is certainly convergence in each step which we now examine. By this we mean that after each update the weight vector moves closer to the optimal vector. From Eq. (2.7) it is obvious that for $t > 0$ we have

$$\bar{u}_t \cdot \bar{u} > 0, \tag{2.12}$$

where $\bar{u}_t$ is the weight vector $\bar{a}_t$ normalised to unity. Because of Eq. (2.12) the criterion for stepwise angle convergence of $\bar{u}_t$ to the optimal $\bar{u}$, namely

$$\bar{u}_{t+1} \cdot \bar{u} - \bar{u}_t \cdot \bar{u} > 0, \tag{2.13}$$

can be equivalently written as

$$(\bar{u}_{t+1} \cdot \bar{u})^2 - (\bar{u}_t \cdot \bar{u})^2 > 0. \tag{2.14}$$

The above inequality motivates us to consider the following quantity

$$\begin{aligned} D \equiv (\bar{u}_{t+1} \cdot \bar{u})^2 - (\bar{u}_t \cdot \bar{u})^2 &= \frac{1}{\|\bar{a}_{t+1}\|^2 \|\bar{a}_t\|^2}\left\{(\bar{a}_{t+1} \cdot \bar{u})^2 \|\bar{a}_t\|^2 - (\bar{a}_t \cdot \bar{u})^2 \|\bar{a}_{t+1}\|^2\right\} \\ = \frac{1}{\|\bar{a}_{t+1}\|^2 \|\bar{a}_t\|^2}&\left\{(\bar{a}_t \cdot \bar{u} + \eta f_t \bar{y}_k \cdot \bar{u})^2 \|\bar{a}_t\|^2 - (\bar{a}_t \cdot \bar{u})^2(\|\bar{a}_t\|^2 + \eta^2 f_t^2 \|\bar{y}_k\|^2 + 2\eta f_t \bar{y}_k \cdot \bar{a}_t)\right\} \\ &= 2\eta f_t \frac{(\bar{a}_t \cdot \bar{u})}{\|\bar{a}_{t+1}\|^2}\left\{\bar{y}_k \cdot \bar{u} - (\bar{u}_t \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t) - \frac{\eta f_t}{2(\bar{a}_t \cdot \bar{u})}\left(\|\bar{y}_k\|^2 (\bar{u}_t \cdot \bar{u})^2 - (\bar{y}_k \cdot \bar{u})^2\right)\right\}. \end{aligned} \tag{2.15}$$

Here use has been made of the update rule (2.1). The demand for positivity of D satisfies our objective for stepwise convergence. We observe that $\bar{y}_k \cdot \bar{u}$ appearing in Eq. (2.15) is definitely positive due to Eq. (2.3). Unfortunately, we cannot make the same assertion regarding the other two terms in brackets. However, as the number of steps increases $(\bar{a}_t \cdot \bar{u})$ increases with time as well because of Eq. (2.7), thereby making the term quadratic in η negligible. Moreover, a slight transformation of Eq. (2.2) to

$$\bar{y}_k \cdot \bar{u}_t \leq \frac{b}{\|\bar{a}_t\|} \tag{2.16}$$

shows that the misclassification condition becomes less restrictive with time. As a result the term $(\bar{u}_t \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t)$ keeps decreasing. Thus, for time t larger than a critical time t_c

positivity of D is accomplished. By placing bounds on the terms in brackets in Eq. (2.15) and using Eqs. (2.2), (2.3) and (2.7) we obtain

$$\begin{aligned}\bar{y}_k \cdot \bar{u} - (\bar{u}_t \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t) - \frac{\eta f_t}{2(\bar{a}_t \cdot \bar{u})} \left(\|\bar{y}_k\|^2 (\bar{u}_t \cdot \bar{u})^2 - (\bar{y}_k \cdot \bar{u})^2 \right) &\geq \gamma_d - \frac{b}{\|\bar{a}_t\|} - \frac{\eta f_{max}}{2(\bar{a}_t \cdot \bar{u})} \left(R^2 - \gamma_d^2 \right) \\ &\geq \gamma_d - \frac{1}{2\eta f_{min} \gamma_d t} \left(2b + \eta f_{max} (R^2 - \gamma_d^2) \right).\end{aligned} \tag{2.17}$$

From the above inequality and demanding positivity of D the time sufficient for stepwise convergence to begin is

$$t_c \equiv \frac{1}{2} \frac{f_{max}}{f_{min}} \frac{R^2}{\gamma_d^2} \left(1 - \frac{\gamma_d^2}{R^2} \right) + \frac{1}{\eta f_{min}} \frac{b}{\gamma_d^2}. \tag{2.18}$$

Between t_c and the time t_N, derived from Novikoff's demand that the algorithm converges eventually, the following inequality holds

$$t_N > 2 \frac{f_{max}}{f_{min}} t_c. \tag{2.19}$$

Therefore, unless the algorithm terminates much before Novikoff's time bound is exhausted, it will definitely enter the phase of stepwise convergence.

It would be interesting to estimate the margin that the algorithm is able to achieve. By substituting Novikoff's time t_N into Eq. (2.9) we obtain a time-independent upper bound on $\|\bar{a}_t\|$

$$\|\bar{a}_t\| \leq \frac{\eta R^2 + 2b}{\gamma_d} \tag{2.20}$$

which, in turn, provides a lower bound β_{min} on the directional margin $\beta = \frac{b}{\|\bar{a}_t\|}$ appearing in the misclassification condition of Eq. (2.16)

$$\beta_{min} = \frac{f_{min}}{f_{max}} \frac{\gamma_d}{\left(2 + \eta f_{max} \frac{R^2}{b} \right)}. \tag{2.21}$$

We see that the maximal guaranteed value of the directional margin that the algorithm is able to achieve is $\frac{1}{2} \frac{f_{min}}{f_{max}} \gamma_d$ for vanishingly small values of the learning rate η or for $b \gg R^2$ [5]. Notice that the existence of a directional margin means that there exists a geometric margin at least as large as the directional one. This is due to the fact that the projection of the augmented weight vector $\bar{a}_t$ onto the original space has a length which cannot exceed $\|\bar{a}_t\|$.

In our analysis so far we required that the function f_t appearing in the update rule of Eq. (2.1) be bounded as in Eq. (2.5) in order for the algorithm to converge. However,

although a positive and bounded f_t is a sufficient condition for convergence it is by no means a necessary one. To illustrate the above statement we consider the function

$$f_t = \frac{b_u - \bar{a}_t \cdot \bar{y}_k}{\|\bar{y}_k\|^2} \tag{2.22}$$

with b_u even slightly larger than the parameter b of the misclassification condition of Eq. (2.2). This update is a minor modification of the well-known single-sample relaxation algorithm with margin [4] in which $b_u = b$ so that f_t is allowed to vanish. We observe that

$$f_{min} = \frac{b_u - b}{R^2} > 0 \tag{2.23}$$

leading to a lower bound on $\|\bar{a}_t\|$ as in Eq. (2.7). In contrast, no upper bound f_{max} exists since f_t can increase indefinitely if $\bar{a}_t \cdot \bar{y}_k$ is negative and large. Nevertheless we can obtain an upper bound on $\|\bar{a}_t\|$ as we shall see shortly. To this end we calculate the difference of the squared norms of the weight vectors in consecutive steps

$$\|\bar{a}_{t+1}\|^2 - \|\bar{a}_t\|^2 = 2\eta(2-\eta)\frac{b_u - \bar{a}_t \cdot \bar{y}_k}{\|\bar{y}_k\|^2}\left\{\frac{b_u}{2-\eta} - \frac{1}{2}(b_u - \bar{a}_t \cdot \bar{y}_k)\right\} \tag{2.24}$$

and we notice that the r.h.s. of the above equation has a maximum with respect to the quantity $(b_u - \bar{a}_t \cdot \bar{y}_k)$ for

$$(b_u - \bar{a}_t \cdot \bar{y}_k)_{opt} = \frac{b_u}{2-\eta}, \tag{2.25}$$

provided $0 < \eta < 2$. Substituting this value in Eq. (2.24) we obtain

$$\|\bar{a}_{t+1}\|^2 - \|\bar{a}_t\|^2 \leq \frac{\eta}{(2-\eta)}\frac{b_u^2}{r^2} \tag{2.26}$$

where $r = \min_k \|\bar{y}_k\|$. Then, a repeated application of the above inequality leads to the upper bound

$$\|\bar{a}_t\|^2 \leq \frac{\eta}{(2-\eta)}\frac{b_u^2}{r^2}t. \tag{2.27}$$

Combining Eqs. (2.7) and (2.27) we get the squeezing relationship

$$\eta f_{min}\gamma_d t \leq \bar{a}_t \cdot \bar{u} \leq \|\bar{a}_t\| \leq \frac{b_u}{r}\sqrt{\frac{\eta}{2-\eta}t} \tag{2.28}$$

from which the following time bound for convergence is derived

$$t \leq \frac{1}{\eta(2-\eta)}\left(\frac{R^2}{b_u - b}\right)^2 \frac{b_u^2}{r^2{\gamma_d}^2}. \tag{2.29}$$

2.2 Algorithms with fixed directional margin condition

Next we examine algorithms where the misclassification condition assumes the form

$$\bar{y}_k \cdot \bar{u}_t \leq \beta, \tag{2.30}$$

where β is a positive parameter. Notice that the above condition amounts to requiring a minimum directional margin which is not lowered with the number of steps. Therefore, successful termination of the algorithm leads to a solution with a guaranteed geometric margin at least as large as the directional margin β found. This is an important difference from the misclassification condition of Eq. (2.2) which, as Eq. (2.16) illustrates, cannot by itself guarantee a minimum directional margin and consequently a geometric one. The condition for optimally correct classification remains the same as in the previous case

$$\bar{y}_k \cdot \bar{u} \geq \gamma_d > \beta \tag{2.31}$$

while the demand for a positive and bounded f_t according to Eq. (2.5) still holds. As an example of such a bounded function, in addition to the commonly used $f_t = 1$, we mention the function

$$f_t = 1 - \beta \frac{\bar{u}_t \cdot \bar{y}_k}{\|\bar{y}_k\|^2}. \tag{2.32}$$

We consider two cases depending on whether the length of the weight vector is free or fixed.

2.2.1 Algorithms with free-length weight vector

In the usual case that the weight vector is free to grow indefinitely a repeated application of Eq. (2.6) with the assumption of initialisation of $\bar{a}_t$ from zero leads again to Eq. (2.7). As a concequence Eq. (2.12) is once more recovered. Therefore, positivity of D is equivalent to stepwise convergence. Placing a lower bound on the term of D which is linear in η we obtain

$$\bar{y}_k \cdot \bar{u} - (\bar{u}_t \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t) \geq \gamma_d - \beta, \tag{2.33}$$

which is definitely positive on account of Eq. (2.31). Furthermore, because of Eq. (2.7) the terms quadratic in η which are not necessarily positive become less important with time leading to positivity of D for t larger than a critical time t_c. Using Eqs. (2.7), (2.30) and (2.31) we can place a constant lower bound on the quantity in D appearing in brackets, i.e.

$$\bar{y}_k \cdot \bar{u} - (\bar{u}_t \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t) - \frac{\eta f_t}{2(\bar{a}_t \cdot \bar{u})} \left(\|\bar{y}_k\|^2 (\bar{u}_t \cdot \bar{u})^2 - (\bar{y}_k \cdot \bar{u})^2 \right) \geq \gamma_d - \beta - \frac{1}{2} \frac{f_{max}}{f_{min}} \frac{1}{\gamma_d t} (R^2 - \gamma_d^2). \tag{2.34}$$

From the above inequality and the requirement that D be positive the estimated time sufficient for the onset of stepwise convergence is

$$t_c \equiv \frac{1}{2}\frac{f_{max}}{f_{min}}\frac{R^2}{\gamma_d^2}\frac{\left(1-\frac{\gamma_d^2}{R^2}\right)}{\left(1-\frac{\beta}{\gamma_d}\right)}. \tag{2.35}$$

It is worth noticing that the critical time t_c turns out to be independent of the learning rate η.

Now that we have guaranteed the convergence of the algorithm as a consequence of the stronger statement of stepwise convergence we proceed to a derivation of a time bound. Our procedure will be to provide an upper bound on $\|\bar{a}_t\|$ which together with the lower one of Eq. (2.7) are finally combined in a Novikoff-like squeezing relationship. For the derivation of an upper bound we first use Eq. (2.1) to obtain

$$\|\bar{a}_{t+1}\|^2 = \|\bar{a}_t\|^2 + 2\eta f_t \bar{y}_k \cdot \bar{a}_t + \eta^2 f_t^2 \|\bar{y}_k\|^2 = \|\bar{a}_t\|^2 \left(1 + \frac{2\eta f_t}{\|\bar{a}_t\|}\bar{y}_k \cdot \bar{u}_t + \frac{\eta^2 f_t^2}{\|\bar{a}_t\|^2}\|\bar{y}_k\|^2\right). \tag{2.36}$$

Taking the square root and using the inequality $\sqrt{1+x} \leq 1 + \frac{x}{2}$ we have

$$\|\bar{a}_{t+1}\| \leq \|\bar{a}_t\| \left(1 + \frac{\eta f_t}{\|\bar{a}_t\|}\bar{y}_k \cdot \bar{u}_t + \frac{1}{2}\frac{\eta^2 {f_t}^2}{\|\bar{a}_t\|^2}\|\bar{y}_k\|^2\right). \tag{2.37}$$

We now observe that the difference of $\|\bar{a}_t\|$ at successive time instants satisfies the inequality

$$\|\bar{a}_{t+1}\| - \|\bar{a}_t\| \leq \eta f_{max}\beta + \frac{\eta}{2}\frac{f_{max}^2}{f_{min}}\frac{R^2}{\gamma_d}\frac{1}{t}. \tag{2.38}$$

Here we have made use of the lower bound on $\|\bar{a}_t\|$ given by Eq. (2.7) and of the misclassification condition of Eq. (2.30). A repeated application of the above inequality $t - N$ times gives

$$\|\bar{a}_t\| - \|\bar{a}_N\| \leq \eta f_{max}\beta(t-N) + \frac{\eta}{2}\frac{f_{max}^2}{f_{min}}\frac{R^2}{\gamma_d}\left(\frac{1}{N} + \frac{1}{N+1} + \ldots + \frac{1}{t-1}\right). \tag{2.39}$$

Since we initialise the weight vector from zero $\|\bar{a}_N\|$, which is entirely generated by the first N updates, satisfies the obvious bound

$$\|\bar{a}_N\| \leq \eta f_{max} R N. \tag{2.40}$$

Replacing $\|\bar{a}_N\|$ by this upper bound into Eq. (2.39) and employing the inequality

$$\sum_{k=n_1}^{n_2}\frac{1}{k} \leq \int_{n_1}^{n_2}\frac{dt}{t} + \frac{1}{n_1} = \ln\frac{n_2}{n_1} + \frac{1}{n_1}, \tag{2.41}$$

justified by the fact that $\frac{1}{t}$ decreases monotonically, we obtain the following upper bound on $\|\bar{a}_t\|$

$$\|\bar{a}_t\| \leq \eta f_{max} \left\{ RN + \beta(t-N) + \frac{1}{2}\frac{f_{max}}{f_{min}}\frac{R^2}{\gamma_d}\left(\ln\frac{t-1}{N} + \frac{1}{N}\right)\right\}. \tag{2.42}$$

Squeezing $\|\bar{a}_t\|$ between its lower bound of Eq. (2.7) and its upper bound of Eq. (2.42) we obtain a relation which constrains the growth of the number t of the steps of the algorithm

$$\left\{ N\frac{f_{min}}{f_{max}}\frac{\gamma_d}{R}\left(1-\frac{f_{min}}{f_{max}}\frac{\gamma_d}{R}\right) + \frac{1}{2N} + \ln\sqrt{\frac{t-1}{N}}\right\}^{-1}(t-N) \leq \left(\frac{f_{max}}{f_{min}}\frac{R}{\gamma_d}\right)^2 \frac{1}{\left(1-\frac{f_{max}}{f_{min}}\frac{\beta}{\gamma_d}\right)}. \tag{2.43}$$

Taking $N=1$ and noticing that

$$\frac{f_{min}}{f_{max}}\frac{\gamma_d}{R}\left(1-\frac{f_{min}}{f_{max}}\frac{\gamma_d}{R}\right) \leq \frac{1}{4}, \tag{2.44}$$

since the function $x(1-x)$ has a maximum value of $\frac{1}{4}$, we obtain the looser but simpler-looking bound

$$\frac{t-1}{3+\ln(t-1)^2} \leq \frac{1}{4}\left(\frac{f_{max}}{f_{min}}\frac{R}{\gamma_d}\right)^2 \frac{1}{\left(1-\frac{f_{max}}{f_{min}}\frac{\beta}{\gamma_d}\right)}. \tag{2.45}$$

Minimising the upper bound of Eq. (2.42) with respect to N we obtain the optimal value

$$N_{opt} = \left[\frac{1}{2}\frac{f_{max}}{f_{min}}\frac{R}{\gamma_d}\frac{1}{\left(1-\frac{\beta}{R}\right)}\right] + 1, \tag{2.46}$$

where $[x]$ denotes the integer part of x. For the near-optimal choice $N = N_{opt}-1$, assuming $N_{opt} > 1$, and noticing that

$$(N_{opt}-1)\frac{f_{min}}{f_{max}}\frac{\gamma_d}{R}\left(1-\frac{f_{min}}{f_{max}}\frac{\gamma_d}{R}\right) \leq \frac{1}{2} \tag{2.47}$$

we obtain the bound

$$\frac{t-N_{opt}+1}{1+\frac{1}{N_{opt}-1}+\ln\left(\frac{t-1}{N_{opt}-1}\right)} \leq \frac{1}{2}\left(\frac{f_{max}}{f_{min}}\frac{R}{\gamma_d}\right)^2 \frac{1}{\left(1-\frac{f_{max}}{f_{min}}\frac{\beta}{\gamma_d}\right)}. \tag{2.48}$$

We would like to point out that unless $f_{min}\gamma_d - f_{max}\beta$ is positive the inequalities (2.43), (2.45) and (2.48) do not lead to upper bounds on t. However, this failure of obtaining an upper bound on the number of steps does not reflect lack of convergence which has already been proved independently. Actually, convergence occurs in a finite number of steps given that $\|\bar{a}_t\|$ increases at most linearly with time. Of course, for the perceptron-like algorithm of this type where $f_t = 1$ we have an upper bound in all cases which interestingly enough

has a dependence on the difference between the optimal directional margin γ_d and the input directional margin β that the algorithm is seeking. The same difference appears in the expression for the critical time t_c of Eq. (2.35). Another extremely interesting property of all algorithms of the class we are discussing is the independence of the time bound on the learning rate η a property shared by the perceptron algorithm with zero margin. This independence from the learning rate has already been apparent from Eq. (2.35) giving the time for the onset of stepwise convergence.

2.2.2 Algorithms with fixed-length weight vector

Finally we examine a class of algorithms in which a condition identical to that of Eq. (2.30) is checked in order to decide whether a training pattern is characterised as misclassified. The main difference with respect to the previous case is that the augmented weight vector has constant length throughout the algorithm. This is achieved by a renormalisation of the length of the newly produced weight vector to the value β defined in Eq. (2.30) each time the update of Eq. (2.1) takes place, i.e.

$$\bar{a}_{t+1} = \beta \frac{\bar{a}_{t+1}}{\|\bar{a}_{t+1}\|} = \beta \bar{u}_{t+1}. \tag{2.49}$$

Like in the previous algorithms we demand that $\bar{u}_t \cdot \bar{u} > 0$ for all t. This condition is ensured by an appropriate choice of the initial condition. Notice that in this particular class of algorithms $\bar{a}_t$ cannot be initialised from zero since use of the unit vector $\bar{u}_t$ is made in each update. We propose that the initial unit vector $\bar{u}_0$ be chosen in the direction of one of the $\bar{y}_k$'s. In this case, due to the form of the update rule and the positivity of f_t, it is obvious that the vector $\bar{a}_t$ is a linear combination with positive coefficients of the training patterns. Therefore, since according to Eq. (2.31) $\bar{y}_k$ satisfies $\bar{y}_k \cdot \bar{u} > 0$ the same is true for $\bar{a}_t$ and consequently for $\bar{u}_t$. Positivity of $\bar{u}_t \cdot \bar{u}$ allows us to use positivity of D defined by Eq. (2.15) as a criterion for stepwise convergence. Taking a closer look at D reveals that according to Eq. (2.33) the term linear in η remains positive throughout the algorithm. For the term quadratic in η which has no definite sign we conclude that an appropriate choice of η can render it smaller than the term linear in η, thereby leading to stepwise convergence from the first step of the algorithm. More specifically, by placing lower bounds on the quantity appearing in brackets in Eq. (2.15) using Eqs. (2.30), (2.31) and (2.49) we have

$$\bar{y}_k \cdot \bar{u} - (\bar{u}_t \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t) - \frac{\eta f_t}{2\,\|\bar{a}_t\|} \left(\|\bar{y}_k\|^2 (\bar{u}_t \cdot \bar{u}) - \frac{(\bar{y}_k \cdot \bar{u})^2}{\bar{u}_t \cdot \bar{u}} \right) \geq \gamma_d - \beta - \frac{\eta f_{max}}{2\beta}(R^2 - \gamma_d^2). \tag{2.50}$$

Positivity of D is achieved for values of η smaller than the critical value η_c

$$\eta_c \equiv \frac{2}{f_{max}} \frac{(\gamma_d - \beta)\beta}{(R^2 - \gamma_d^2)}. \tag{2.51}$$

After having shown that the algorithm converges step by step our next move will be to place an upper bound on the number of the updates. We proceed by placing a lower bound on $\frac{1}{\|\bar{a}_{t+1}\|}$ employing Eq. (2.37) and the inequality $(1+x)^{-1} \geq 1 - x$

$$\frac{1}{\|\bar{a}_{t+1}\|} \geq \frac{1}{\|\bar{a}_t\|} \left(1 - \frac{\eta f_t}{\|\bar{a}_t\|} \bar{y}_k \cdot \bar{u}_t - \frac{1}{2} \frac{\eta^2 {f_t}^2}{\|\bar{a}_t\|^2} \|\bar{y}_k\|^2 \right). \tag{2.52}$$

Using the above inequality and the update rule we have

$$\bar{u}_{t+1} \cdot \bar{u} = \frac{\bar{a}_t \cdot \bar{u} + \eta f_t \bar{y}_k \cdot \bar{u}}{\|\bar{a}_{t+1}\|} \geq (\bar{u}_t \cdot \bar{u} + \frac{\eta f_t}{\|\bar{a}_t\|} \bar{y}_k \cdot \bar{u}) \left(1 - \frac{\eta f_t}{\|\bar{a}_t\|} \bar{y}_k \cdot \bar{u}_t - \frac{1}{2} \frac{\eta^2 {f_t}^2}{\|\bar{a}_t\|^2} \|\bar{y}_k\|^2 \right) = \bar{u}_t \cdot \bar{u}$$
$$+ \frac{\eta f_t}{\|\bar{a}_t\|} \left\{ \bar{y}_k \cdot \bar{u} - (\bar{u}_t \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t) - \frac{1}{2} \frac{\eta f_t}{\|\bar{a}_t\|} \left(\|\bar{y}_k\|^2 \bar{u}_t \cdot \bar{u} + 2(\bar{y}_k \cdot \bar{u})(\bar{y}_k \cdot \bar{u}_t) \right) - \frac{1}{2} \frac{\eta^2 {f_t}^2}{\|\bar{a}_t\|^2} \|\bar{y}_k\|^2 \bar{y}_k \cdot \bar{u} \right\}. \tag{2.53}$$

We now observe that the difference $\bar{u}_{t+1} \cdot \bar{u} - \bar{u}_t \cdot \bar{u}$ can be bounded from below by a constant

$$\bar{u}_{t+1} \cdot \bar{u} - \bar{u}_t \cdot \bar{u} \geq \frac{\eta f_{min}}{\beta} \left\{ (\gamma_d - \beta) - \frac{\eta f_{max}}{2\beta} R^2 \left(1 + \frac{2\beta}{R} \right) - \frac{\eta^2 f_{max}^2}{2\beta^2} R^3 \right\}. \tag{2.54}$$

Here we made use of Eqs. (2.30), (2.31) and of the fact that $\|\bar{a}_t\| = \beta$. A repeated application of Eq. (2.54) with a rearrangement of the terms on its r.h.s. in powers of $\left(\frac{\eta f_{max} R}{\beta} \right)$ gives

$$\bar{u}_t \cdot \bar{u} - \bar{u}_0 \cdot \bar{u} \geq \frac{f_{min}}{f_{max}} \left\{ \frac{\gamma_d - \beta}{R} \left(\frac{\eta f_{max} R}{\beta} \right) - \frac{1}{2} \left(1 + \frac{2\beta}{R} \right) \left(\frac{\eta f_{max} R}{\beta} \right)^2 - \frac{1}{2} \left(\frac{\eta f_{max} R}{\beta} \right)^3 \right\} t. \tag{2.55}$$

By setting the final condition $\bar{u}_t \cdot \bar{u} = 1$ implying convergence and taking into account that $\bar{u}_0 \cdot \bar{u} > 0$ we obtain the time bound

$$t < \frac{f_{max}}{f_{min}} \left\{ \frac{\gamma_d - \beta}{R} \left(\frac{\eta f_{max} R}{\beta} \right) - \frac{1}{2} \left(1 + \frac{2\beta}{R} \right) \left(\frac{\eta f_{max} R}{\beta} \right)^2 - \frac{1}{2} \left(\frac{\eta f_{max} R}{\beta} \right)^3 \right\}^{-1}. \tag{2.56}$$

The above time bound can be optimised with respect to the parameter η. The resulting optimal value of η is approximately given by

$$\eta_{opt} = \frac{1}{f_{max}} \frac{(\gamma_d - \beta)\beta}{R^2} \left(1 + \frac{2\beta}{R} \right)^{-1}. \tag{2.57}$$

Substituting the optimal value of η into Eq. (2.56) we obtain the optimised time bound

$$t < 2\frac{f_{max}}{f_{min}}\frac{R^2}{(\gamma_d-\beta)^2}\left(1+\frac{2\beta}{R}\right)\left(1-\frac{\gamma_d-\beta}{R}\left(1+\frac{2\beta}{R}\right)^{-2}\right)^{-1}. \tag{2.58}$$

From the above expression we observe that our time bound is analogous to the one of the perceptron without margin with the main differences being a factor of 2 and the replacement of γ_d^2 by $(\gamma_d-\beta)^2$.

2.3 Estimating the optimal geometric margin

In this subsection we attempt to place an upper bound on the optimal geometric margin of a training set in terms of the optimal directional margin in an augmented space which is the original one supplemented with an additional dimension. All the patterns are placed in the position $\rho_0 = \rho > 0$ in that additional dimension and then a reflection with respect to the origin is performed. As a result of such a reflection the patterns that fall into the first category (positive projection on the weight vector) have the coordinate $\rho_0 = \rho$ in the additional dimension with the others (negative projection on the weight vector) having the coordinate $\rho_0 = -\rho$.

If we denote by $\bar{a} = [\bar{w}\ w_0]$ a weight vector in the augmented space that classifies the patterns correctly then the geometric margin $\gamma(\bar{a})$ of the set can be calculated from

$$\gamma(\bar{a}) = \min_k \{\bar{a}\cdot\bar{y}_k\}\,\|\bar{w}\|^{-1} = \min_k \left\{[\bar{w}\ w_0][\bar{x}_k\ \rho_0]^T\right\}\|\bar{w}\|^{-1} = \min_k \{\bar{w}\cdot\bar{x}_k + w_0\rho_0\}\,\|\bar{w}\|^{-1}, \tag{2.59}$$

where $\bar{w}$ and $\bar{x}_k$ are the components in the original space of $\bar{a}$ and $\bar{y}_k$, respectively and $\frac{|w_0|}{\|\bar{w}\|}\rho$ is the distance from the origin of the hyperplane normal to $\bar{w}$. Since the maximum value that this distance can take is $R_x = \max_k \|\bar{x}_k\|$ we obtain

$$\frac{|w_0|}{\|\bar{w}\|} \leq \frac{R_x}{\rho}. \tag{2.60}$$

The directional margin $\gamma_d(\bar{a})$ that corresponds to $\gamma(\bar{a})$ can be evaluated using the relationship

$$\gamma_d(\bar{a}) = \frac{\|\bar{w}\|}{\|\bar{a}\|}\gamma(\bar{a}) \tag{2.61}$$

from which

$$\gamma_d(\bar{a}) \leq \gamma(\bar{a}) \tag{2.62}$$

follows since $\|\bar{w}\| \leq \|\bar{a}\|$. Taking the norm of $\bar{a}$ we obtain

$$\|\bar{a}\| = \sqrt{\|\bar{w}\|^2 + w_0^2} \leq \|\bar{w}\| \sqrt{1 + \frac{R_x^2}{\rho^2}} = \|\bar{w}\| \frac{R}{\rho}. \tag{2.63}$$

Here use has been made of Eq. (2.60) and of the fact that

$$R^2 = \rho^2 + R_x^2. \tag{2.64}$$

Substituting Eq. (2.63) in Eq. (2.61) we get

$$\gamma(\bar{a}) \leq \frac{R}{\rho} \gamma_d(\bar{a}). \tag{2.65}$$

In the case that the weight vector $\bar{a}$ is the optimal one $\bar{a}_{opt}$ maximising the geometric margin we have

$$\gamma \equiv \gamma(\bar{a}_{opt}) \leq \frac{R}{\rho} \gamma_d(\bar{a}_{opt}). \tag{2.66}$$

Taking into account that $\gamma_d = \max_{\bar{a}} \gamma_d(\bar{a}) \geq \gamma_d(\bar{a}_{opt})$ and $\gamma = \max_{\bar{a}} \gamma(\bar{a}) \geq \max_{\bar{a}} \gamma_d(\bar{a}) = \gamma_d$ the above inequality leads to

$$1 \leq \frac{\gamma}{\gamma_d} \leq \frac{R}{\rho}. \tag{2.67}$$

For $\rho = 1$ Eq. (2.67) gives $\gamma \leq R\gamma_d$. By placing the patterns at a distance R_x in the additional dimension we achieve an optimal geometric margin of at most $\sqrt{2}\gamma_d$ [3]. In the limit $\rho \to \infty$ Eq. (2.64) implies that $\frac{R}{\rho} \to 1$. Then from Eq. (2.67) follows that in this limit the optimal directional margin γ_d tends to the optimal geometric one γ

$$\lim_{\rho \to \infty} \gamma_d = \gamma. \tag{2.68}$$

The above analysis leads to the important conclusion that an algorithm seeking the optimal directional margin is equivalent to an algorithm that looks for the optimal geometric margin if the training patterns are translated infinitely far from the origin in the augmented space. This, of course, is achieved at an infinite computational cost since R, which appears in the time bounds, tends to infinity.

3 Algorithmic implementation

In this section we present two algorithms seeking the optimal directional margin which, however, due to the analysis of subsection 2.3 could be used to approximately obtain the optimal geometric margin. The data used by both algorithms are mapped into an

augmented space where the length of the translation $|\rho|$ in the additional dimension is treated as a free parameter controlling the balance between the geometric margin to be achieved and the computational cost.

The first implementation makes repeated use of the free-length weight vector algorithm of subsection 2.2.1 with any positive and bounded function f_t in its update rule. The choice of this specific algorithm is justified by the fact that it is independent of the learning rate η which otherwise would have to be appropriately tuned. In each round of its repeated application the algorithm looks for a fixed unrelaxed directional margin β according to the condition $\bar{a}_t \cdot \bar{y}_k > \beta$. Each round lasts until the condition is satisfied by all the training patterns or until an upper bound on the number of checks over the training set is reached. The range of values that β can take and therefore the interval that the algorithm should search extends from 0 to $r = \min_k \|\bar{y}_k\|$. The search can be performed efficiently by using a procedure similar to the Bolzano-bisection method. Initially $\bar{a}_0$ is set to zero and a margin $\beta = \frac{r}{2}$ is asked for with a step parameter being set to $\frac{r}{2}$. If the algorithm comes up with a solution vector $\bar{a}$ satisfying the imposed margin constraint without exhausting the upper number of checks the round is considered successful. The weight vector $\bar{a}$ is stored as the best solution found so far and is exploited as the initial value $\bar{a}_0$ of the next trial. This way the procedure of finding a better solution in a subsequent round is speeded up substantially since such an $\bar{a}$ lies probably closer to a weight vector which gives rise to a larger margin than the weight vector $\bar{a}_0 = 0$ and thus constitutes a better guess as an initial condition. At the end of each trial the step is divided by 2. In the case that a trial ends successfully the target value of the margin β in the next round is calculated by adding to the previous one the present step otherwise β is reduced by the same amount. Therefore, on the condition that the upper number of checks is set to a sufficiently large value, the procedure guarantees that the deviation of the margin β from the maximum one is reduced by a factor of 2 in each round. The algorithm is terminated when the step reaches a certain predefined desirable level, thereby determining dynamically the number of rounds.

The second implementation tries to take advantage to some extent of the time spent in unsuccessful trials.To accomplish this the upper bound on the number of checks of the condition in each round is divided between a module with the condition mentioned above and one which uses the relaxed condition $\bar{a}_t \cdot \bar{y}_k > b$. If the number of checks dedicated to the first module is exhausted without the condition being satisfied by all the patterns then the algorithm proceeds to the second module. There, in the place of b we use $\beta \|\bar{a}_f\|$ where $\bar{a}_f$ is the $\bar{a}_t$ when leaving the first module. If the second module terminates without exhausting the specified number of checks the directional margin is computed as $\beta \|\bar{a}_f\| / \|\bar{a}_s\|$, where $\bar{a}_s$ is the weight vector when leaving the second module. The round is considered successful

only if a solution is found during the execution of the first module. At the end of each trial the margin found is compared to the best one until that point and the largest of the two is kept together with the solution weight vector found which is exploited as an initial condition of the next round. The change of β in each round is performed here the same way as in the first implementation. The only difference is that if β as calculated from the bisection procedure is less than the best margin stored (obviously as a result of a successful second module of some previous unsuccessful trial) then this value of β is already achieved as a margin by the training patterns and therefore the algorithm proceeds without checking the misclassification condition considering the present step as successful.

Before concluding this section we would like to emphasise that, although the time required by our algorithmic procedure to find a near-optimum margin is not necessarily smaller than the time required by other methods if the training sets are relatively small, our method is certainly faster for large data sets. This is due to the fact that our algorithm, which does not use dual variables, has a running time which scales linearly with the number of training points.

4 Conclusions

In summary, we examined the convergence of perceptron-like algorithms with margin and developed a criterion for the stronger requirement of stepwise convergence which allowed us to adopt a unified approach in the theoretical analysis. We also proposed a new class of such algorithms in which the standard classification condition is replaced by a more stringent one which insits on a fixed value of the directional margin and proved that they converge in a finite number of steps. An algorithmic implementation reminiscent of Bolzano's bisection method made possible a fast search through the whole interval of allowed values for the optimal directional margin. We subsequently showed that as the distance in which the data are placed in the additional dimension of the augmented space increases the optimal directional margin approaches the optimal geometric one. This observation transforms our algorithmic procedure into a fast and simple approximate maximal margin classifier.

References

[1] M. A. Aizerman, E. M. Braverman and L. I. Rozonoer (1964). 'Theoretical foundations of the potential function method in pattern recognition learning'. *Automation and remote control* **25**, 821-837.

[2] B. E. Boser, I. M. Guyon and V. N. Vapnik (1992). 'A training algorithm for optimal margin classifiers'. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144-152, ACM Press.

[3] N. Cristianini and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines.* Cambridge, UK: Cambridge University Press.

[4] R. O. Duda, P. E. Hart and D. G. Stork (2000). *Pattern Classsification.* Wiley-Interscience, 2nd edition.

[5] W. Krauth and M. Mezard (1987). 'Learning algorithms with optimal stability in neural networks'. *Journal of Physics* **A 20** L745-L752.

[6] A. B. J. Novikoff (1962). 'On convergence proofs on perceptrons'. In *Proceedings of the Symposium on the Mathematical Theory of Automata,* Volume 12, pp. 615-622, Polytechnic Institute of Brooklyn.

[7] F. Rosenblatt (1958). 'The perceptron: A probabilistic model for information storage and organization in the brain'. *Psychological Review* **65**(6), 386-408.

[8] V. Vapnik (1995). *The Nature of Statistical Learning Theory.* Springer Verlag.