Constant Rate Approximate Maximum Margin Algorithms

Petroula Tsampouka and John Shawe-Taylor

ECS, University of Southampton, UK
e-mail: {pt04r, jst}@ecs.soton.ac.uk

Abstract. We present a new class of perceptron-like algorithms with margin in which the "effective" learning rate, defined as the ratio of the learning rate to the length of the weight vector, remains constant. We prove that the new algorithms converge in a finite number of steps and show that there exists a limit of the parameters involved in which convergence leads to classification with maximum margin.

1 Introduction

It is generally believed that the larger the margin of the solution hyperplane the greater is the generalisation ability of the learning machine [11, 9]. The simplest on-line learning algorithm for binary linear classification, Rosenblatt's Perceptron [8], does not aim at any margin. The problem, instead, of finding the optimal margin hyperplane lies at the core of Support Vector Machines (SVMs) [11, 2]. SVMs, however, require solving a quadratic programming problem which makes their efficient implementation difficult and, often, time consuming.

The difficulty in implementing SVMs has respurred a lot of interest in alternative large margin classifiers many of which are based on the Perceptron algorithm. The most well-known such variants are the standard Perceptron with margin [3, 6, 7, 10] and the ALMA [4] algorithms both similar in many respects. Our purpose here is to address the maximum margin classification problem in the context of perceptron-like algorithms which, however, differ from the above mentioned variants in the sense that the learning rate varies with time in more or less the same way as the length of the weight vector does. This new class of algorithms emerged from an attempt to classify perceptron-like classifiers with margin in a few very broad categories according to the dependence on time of the misclassification condition or of the effect that an update has on the current weight vector. The new algorithms are shown to converge in a finite number of steps to an approximation of the optimal solution vector which becomes better as the parameters involved follow a specific limiting process.

A taxonomy of perceptron-like large margin classifiers can be found in Sect. 2. The new algorithm, called Constant Rate Approximate Maximum Margin Algorithm (CRAMMA), is described in Sect. 3 together with an analysis regarding its convergence. Section 4 provides experimental evidence supporting the theoretical analysis put forward in Sect. 3. Finally, Sect. 5 contains our conclusions.

2 Taxonomy of Perceptron-Like Large Margin Classifiers

In what follows we make the assumption that we are given a training set which, even if not initially linearly separable can, by an appropriate feature mapping into a space of a higher dimension [1, 11, 2] be classified into two categories by a linear classifier. This higher dimensional space in which the patterns are linearly separable will be the considered space. By adding one additional dimension and placing all patterns in the same position at a distance ρ in that dimension we construct an embedding of our data into the so-called augmented space [3]. The advantage of this embedding is that the linear hypothesis in the augmented space becomes homogeneous. Thus, only hyperplanes passing through the origin in the augmented space need to be considered even for tasks requiring bias. Throughout our discussion a reflection with respect to the origin in the augmented space of the negatively labelled patterns is assumed in order to allow for a uniform treatment of both categories of patterns. Also, we use the notation $R = \max_k \|\boldsymbol{y}_k\|$ and $r = \min_k \|\boldsymbol{y}_k\|$, where $\boldsymbol{y}_k$ is the k^{th} augmented pattern. Obviously, $R \geq r \geq \rho$.

The relation characterising optimally correct classification of the training patterns $\boldsymbol{y}_k$ by a weight vector $\boldsymbol{u}$ of unit norm in the augmented space is

$$\boldsymbol{u} \cdot \boldsymbol{y}_k \geq \gamma_{\text{d}} \quad \forall k \ . \tag{1}$$

The quantity γ_{d}, which we call the maximum directional margin, is defined as

$$\gamma_{\text{d}} = \max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \min_k \{\boldsymbol{u} \cdot \boldsymbol{y}_k\}$$

and is obviously bounded from above by r. The maximum directional margin determines the maximum distance from the origin in the augmented space of the hyperplane normal to $\boldsymbol{u}$ placing all training patterns on the positive side and coincides with the maximum margin in the augmented space with respect to hyperplanes passing through the origin if no reflection is assumed. In the determination of this hyperplane only the direction of $\boldsymbol{u}$ is exploited with no reference to its projection onto the original space. As a consequence the maximum directional margin is not necessarily realised with the same weight vector that gives rise to the maximum geometric margin γ in the original space. Notice, however, that

$$1 \leq \frac{\gamma}{\gamma_{\text{d}}} \leq \frac{R}{\rho} \ . \tag{2}$$

In the limit $\rho \to \infty$, $R/\rho \to 1$ and from (2) $\gamma_{\text{d}} \to \gamma$ [10]. Thus, with ρ increasing the maximum directional margin γ_{d} approaches the maximum geometric one γ.

We concentrate on algorithms that update the augmented weight vector $\boldsymbol{a}_t$ by adding a suitable positive amount in the direction of the misclassified (according to an appropriate condition) training pattern $\boldsymbol{y}_k$. The general form of such an update rule is

$$\boldsymbol{a}_{t+1} = \frac{\boldsymbol{a}_t + \eta_t f_t \boldsymbol{y}_k}{N_{t+1}} \ , \tag{3}$$

where η_t is the learning rate which could depend explicitly on the number t of updates that took place so far and f_t an implicit function of the current step (update) t, possibly involving the current weight vector $\boldsymbol{a}_t$ and/or the current misclassified pattern $\boldsymbol{y}_k$, which we require to be positive and bounded, i.e.

$$0 < f_{\min} \leq f_t \leq f_{\max} \ . \tag{4}$$

We also allow for the possibility of normalising the newly produced weight vector $\boldsymbol{a}_{t+1}$ to a desirable length through a factor N_{t+1}. For the Perceptron algorithm η_t is constant, $f_t = 1$ and $N_{t+1} = 1$. Each time the predefined misclassification condition is satisfied by a training pattern the algorithm proceeds to the update of the weight vector. We adopt the convention of initialising t from 1.

A sufficiently general form of the misclassification condition is

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq C(t) \ , \tag{5}$$

where $\boldsymbol{u}_t$ is the weight vector $\boldsymbol{a}_t$ normalised to unity and $C(t) > 0$ if we require that the algorithm achieves a positive margin. If $\boldsymbol{a}_1 = \boldsymbol{0}$ we treat the first pattern in the sequence as misclassified. We distinguish two cases depending on whether $C(t)$ is bounded from above by a strictly decreasing function of t which tends to zero or remains bounded from above and below by constants. In the first case the minimum directional margin required by such a condition becomes lower than any fixed value provided t is large enough. Algorithms with such a condition have the advantage of achieving some fraction of the unknown existing margin provided they converge. Examples of such algorithms are the well-known standard Perceptron algorithm with margin [3, 6, 7, 10] and the ALMA$_2$ algorithm [4]. In the standard Perceptron algorithm with margin the misclassification condition takes the form

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq \frac{b}{\|\boldsymbol{a}_t\|} \ , \tag{6}$$

where $c_1(t-1) \leq \|\boldsymbol{a}_t\| \leq c_2\sqrt{t-1}$ with b, c_1, c_2 positive constants. In the ALMA$_2$ algorithm the misclassification condition is

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq \frac{b}{\|\boldsymbol{a}_t\|\sqrt{t}} \ , \tag{7}$$

in which $c_3\sqrt{t-1} \leq \|\boldsymbol{a}_t\| \leq R$ with b, c_3 positive constants (see Appendix A). Notice the striking similarity characterising the behaviour of $C(t)$ in the Perceptron and ALMA$_2$ algorithms. In the second case the condition amounts to requiring a directional margin, assumed to exist, which is not lowered arbitrarily with the number t of updates. In particular, if $C(t)$ is equal to a constant β [10] (5) becomes

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq \beta \tag{8}$$

and successful termination of the algorithm leads to a solution with margin larger than β. Obviously, convergence is not possible unless $\beta < \gamma_{\mathrm{d}}$. In this case an organised search through the range of possible β values is necessary.

An alternative classification of the algorithms with the perceptron-like update rule (3) is according to the dependence on t of the "effective" learning rate

$$\eta_{\mathrm{eff}\,t} \equiv \frac{\eta_t R}{\|\boldsymbol{a}_t\|} \tag{9}$$

which controls the impact that an update has on the current weight vector. More specifically, $\eta_{\mathrm{eff}\,t}$ determines the update of the direction $\boldsymbol{u}_t$

$$\boldsymbol{u}_{t+1} = \frac{\boldsymbol{u}_t + \eta_{\mathrm{eff}\,t} f_t \boldsymbol{y}_k / R}{\|\boldsymbol{u}_t + \eta_{\mathrm{eff}\,t} f_t \boldsymbol{y}_k / R\|} \ . \tag{10}$$

Again we distinguish two cases depending on whether $\eta_{\mathrm{eff}\,t}$ is bounded from above by a strictly decreasing function of t which tends to zero or remains bounded from above and below by constants. We do not consider the case that $\eta_{\mathrm{eff}\,t}$ increases indefinitely with t since, as we will argue below, we do not expect such algorithms to converge in a finite number of steps. In the first category belong the Perceptron algorithm with both the standard misclassification condition (6) and the fixed directional margin one of (8) [10] in which η_t remains constant and $\|\boldsymbol{a}_t\|$ is bounded from below by a positive linear function of t. Also to the same category belongs the ALMA$_2$ algorithm in which η_t decreases as $1/\sqrt{t}$. The similarity of the standard Perceptron with margin and ALMA$_2$ algorithms with respect to the behaviour of $\eta_{\mathrm{eff}\,t}$ is apparent if we consider the bounds obeyed by $\|\boldsymbol{a}_t\|$ in these two cases. Moreover, in both algorithms $\eta_{\mathrm{eff}\,t}$ is proportional to $C(t)$. In the second category belong algorithms with the fixed directional margin condition of (8), $\|\boldsymbol{a}_t\|$ normalised to the target margin value β and fixed learning rate [10].

A very desirable property of an algorithm is certainly progressive convergence at each step meaning that at each update $\boldsymbol{u}_t$ moves closer to the optimal direction $\boldsymbol{u}$. Let us assume that

$$\boldsymbol{u}_t \cdot \boldsymbol{u} > 0 \ . \tag{11}$$

This condition is readily satisfied provided the initial weight vector either vanishes or is chosen in the direction of any of the patterns $\boldsymbol{y}_k$. Indeed, on account of the update rule (3) and the assumption that there exists a positive margin γ_{d} according to (1) the weight vector is always a linear combination with positive coefficients of vectors which possess a positive inner product with the optimal direction $\boldsymbol{u}$. Because of (11) the criterion for stepwise angle convergence [10], namely

$$\Delta \equiv \boldsymbol{u}_{t+1} \cdot \boldsymbol{u} - \boldsymbol{u}_t \cdot \boldsymbol{u} > 0 \ ,$$

can be equivalently expressed as a demand for positivity of D

$$D \equiv (\boldsymbol{u}_{t+1} \cdot \boldsymbol{u})^2 - (\boldsymbol{u}_t \cdot \boldsymbol{u})^2 = 2\eta_{\mathrm{eff}\,t} f_t (\boldsymbol{u}_t \cdot \boldsymbol{u}) \left\| \boldsymbol{u}_t + \eta_{\mathrm{eff}\,t} f_t \frac{\boldsymbol{y_k}}{R} \right\|^{-2} \frac{A}{R} \ ,$$

where use has been made of (10) and A is defined by

$$A \equiv \boldsymbol{y}_k \cdot \boldsymbol{u} - (\boldsymbol{u}_t \cdot \boldsymbol{u})(\boldsymbol{y}_k \cdot \boldsymbol{u}_t) - \frac{\eta_{\mathrm{eff}\,t} f_t}{2R} \left(\|\boldsymbol{y}_k\|^2 (\boldsymbol{u}_t \cdot \boldsymbol{u}) - \frac{(\boldsymbol{y}_k \cdot \boldsymbol{u})^2}{(\boldsymbol{u}_t \cdot \boldsymbol{u})} \right) \ .$$

Positivity of A leads to positivity of D on account of (4) and (11) and consequently to stepwise convergence. Actually, convergence occurs in a finite number of steps provided that after some time A becomes bounded from below by a positive constant and $\eta_{\mathrm{eff}\,t}$ remains bounded by positive constants or decreases indefinitely but not faster than $1/t$. Following this rather unified approach one can examine whether sooner or later an algorithm enters the stage of stepwise convergence and terminates successfully in a finite number of steps [10].

We are now going to argue that algorithms with $\eta_{\mathrm{eff}\,t}$ growing indefinitely are unlikely to converge in a finite number of steps since such a behaviour is incompatible with the onset of stepwise convergence if such a stepwise convergence leads to convergence in a finite number of steps. Indeed, if we assume that for t larger than a critical value t_c the algorithm enters such a stage then sooner or later $\boldsymbol{u}_t \cdot \boldsymbol{u}$ will increase sufficiently such that $\left(\|\boldsymbol{y}_k\|^2 (\boldsymbol{u}_t \cdot \boldsymbol{u}) - (\boldsymbol{y}_k \cdot \boldsymbol{u})^2 / (\boldsymbol{u}_t \cdot \boldsymbol{u}) \right)$ becomes positive. Multiplication of such a positive term with a sufficiently large $\eta_{\mathrm{eff}\,t}$ will then make A negative contradicting our assumption.

It is not difficult to see that (1), (4), (5) and $R \geq \|\boldsymbol{y}_k\|$ lead to

$$A \geq \gamma_{\mathrm{d}} - C(t) - \frac{1}{2} f_{\max} \eta_{\mathrm{eff}\,t} R \ . \tag{12}$$

By requiring that the r.h.s. of (12) be positive we derive a sufficient condition for the onset of stepwise convergence

$$\eta_{\mathrm{eff}\,t} < 2 \frac{\gamma_{\mathrm{d}} - C(t)}{f_{\max} R} \ . \tag{13}$$

The above condition is always eventually satisfied in the case of algorithms with $\eta_{\mathrm{eff}\,t} \to 0$ in the limit $t \to \infty$ like the Perceptron and ALMA$_2$ algorithms. If this is not the case, however, we are forced to suppress $\eta_{\mathrm{eff}\,t}$ sufficiently.

In summary, the misclassification condition and the effective learning rate of a perceptron-like algorithm could, roughly speaking, either be "relaxed" with time or remain practically constant. Thus, we are led to four broad categories of potentially convergent algorithms. Out of these categories the one with condition "relaxed" with time and fixed effective learning rate has not, to the best of our knowledge, been examined before and is the subject of the present work.

3 The Constant Rate Approximate Maximum Margin Algorithm CRAMMA$^{\epsilon}$

We consider algorithms with constant effective learning rate $\eta_{\mathrm{eff}\,t} = \eta_{\mathrm{eff}}$ in which the misclassification condition takes the form of (5) with

$$C(t) = \frac{\beta}{t^{\epsilon}} \ , \tag{14}$$

where β and ϵ are positive constants. We assume that the initial value $\boldsymbol{u}_1$ of $\boldsymbol{u}_t$ is the unit vector in the direction of the first training pattern in order for (11)

Fig. 1. The constant rate approximate maximum margin algorithm CRAMMA$^{\epsilon}$.

Require: A linearly separable augmented training set with reflection assumed $S = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$
Define:
For $k = 1, \ldots, m$
$R = \max_k \|\boldsymbol{y}_k\|$, $\boldsymbol{y}'_k = \boldsymbol{y}_k / R$
Fix: ϵ, η_{eff}, β_1 $(= \beta/R)$
Initialisation:
$t = 1$, $\boldsymbol{u}_1 = \boldsymbol{y}'_1 / \|\boldsymbol{y}'_1\|$

repeat until no update made within the **for** loop
  for $k = 1$ to m **do**
    if $\boldsymbol{u}_t \cdot \boldsymbol{y}'_k \leq \beta_t$ **then**
      $\boldsymbol{u}_{t+1} = \frac{\boldsymbol{u}_t + \eta_{\text{eff}} \boldsymbol{y}'_k}{\|\boldsymbol{u}_t + \eta_{\text{eff}} \boldsymbol{y}'_k\|}$
      $t = t + 1$
      $\beta_t = \beta_1 / t^{\epsilon}$

to hold. We additionally make the choice $f_t = 1$. Since the above $C(t)$ does not depend on $\|\boldsymbol{a}_t\|$ and given that (the update (10) of) $\boldsymbol{u}_t$ depends on $\|\boldsymbol{a}_t\|$ only through η_{eff} the algorithm does not depend separately on η_t and $\|\boldsymbol{a}_t\|$ but only on their ratio i.e. on η_{eff}. From (13) we obtain the constraint

$$\eta_{\text{eff}} < 2\frac{\gamma_{\text{d}}}{R}$$

on the constant effective learning rate η_{eff} in order for the algorithm to eventually enter the stage of stepwise convergence. Obviously, the further η_{eff} is from this upper bound the earlier the stepwise convergence will begin.

Although only η_{eff} plays a role we still prefer to think of it as arising from a weight vector normalised to the constant value β

$$\|\boldsymbol{a}_t\| = \beta$$

and a learning rate having a fixed value as well

$$\eta_t = \eta \ .$$

This is equivalent to normalising the weight vector to the variable margin value $C(t)$ that the algorithm is after assuming at the same time a variable learning rate $\eta_t = \eta / t^{\epsilon}$. Having in mind our earlier comment regarding the meaning of the directional margin in the augmented space the geometric interpretation of such a choice becomes clear: The algorithm is looking for the hyperplane tangent to a hypersphere centered at the origin of the augmented space of radius $\|\boldsymbol{a}_t\|$ equal to the target margin value $C(t)$ which leaves all the augmented (with a reflection assumed) patterns on the positive side. The t-independent value of the learning rate η might also be considered as dependent on (a power of) β, i.e.

$$\eta = \eta_0 \left(\frac{\beta}{R}\right)^{1-\delta} \ ,$$

where η_0, δ are positive constants. Thus, we are led to an effective learning rate which scales with $\frac{\beta}{R}$ like

$$\eta_{\text{eff}} = \eta_0 \left(\frac{\beta}{R}\right)^{-\delta} \ . \tag{15}$$

The above algorithm with constant effective learning rate $\eta_{\rm eff}$ and misclassification condition described by (5) and (14) involving the power ϵ of t will be called the Constant Rate Approximate Maximum Margin Algorithm CRAMMA$^{\epsilon}$ and is presented in Fig. 1. A justification of the qualification of the algorithm as an "Approximate Maximum Margin" one stems from the following theorem.

Theorem 1. *The* CRAMMA$^{\epsilon}$ *algorithm of Fig. 1 converges in a finite number of steps provided* $\eta_{\rm eff} < \frac{\gamma_{\rm d}}{R}$. *Moreover, if* $\eta_{\rm eff}$ *is given a dependence on* β *through the relation* $\eta_{\rm eff} = \eta_0 \left(\frac{\beta}{R}\right)^{-\delta}$ *the guaranteed fraction of the maximum directional margin* $\gamma_{\rm d}$ *achieved in the limit* $\frac{\beta}{R} \to \infty$ *tends to* 1 *provided* $0 < \epsilon\delta < 1$.

Proof. Taking the inner product of (10) with the optimal direction $\boldsymbol{u}$ we have

$$\boldsymbol{u}_{t+1} \cdot \boldsymbol{u} = \left(\boldsymbol{u}_t \cdot \boldsymbol{u} + \eta_{\rm eff} \frac{\boldsymbol{y}_k \cdot \boldsymbol{u}}{R}\right) \left\| \boldsymbol{u}_t + \eta_{\rm eff} \frac{\boldsymbol{y}_k}{R} \right\|^{-1} \quad . \tag{16}$$

Here

$$\left\| \boldsymbol{u}_t + \eta_{\rm eff} \frac{\boldsymbol{y}_k}{R} \right\|^{-1} = \left(1 + 2\eta_{\rm eff} \frac{\boldsymbol{y}_k \cdot \boldsymbol{u}_t}{R} + \eta_{\rm eff}^2 \frac{\|\boldsymbol{y}_k\|^2}{R^2}\right)^{-\frac{1}{2}}$$

from where, by using the inequality $(1+x)^{-\frac{1}{2}} \geq 1 - \frac{x}{2}$, we get

$$\left\| \boldsymbol{u}_t + \eta_{\rm eff} \frac{\boldsymbol{y}_k}{R} \right\|^{-1} \geq 1 - \eta_{\rm eff} \frac{\boldsymbol{y}_k \cdot \boldsymbol{u}_t}{R} - \eta_{\rm eff}^2 \frac{\|\boldsymbol{y}_k\|^2}{2R^2} \quad .$$

Then, (16) becomes

$$\boldsymbol{u}_{t+1} \cdot \boldsymbol{u} \geq \left(\boldsymbol{u}_t \cdot \boldsymbol{u} + \eta_{\rm eff} \frac{\boldsymbol{y}_k \cdot \boldsymbol{u}}{R}\right) \left(1 - \eta_{\rm eff} \frac{\boldsymbol{y}_k \cdot \boldsymbol{u}_t}{R} - \eta_{\rm eff}^2 \frac{\|\boldsymbol{y}_k\|^2}{2R^2}\right) \quad .$$

Thus, we obtain for $\Delta = \boldsymbol{u}_{t+1} \cdot \boldsymbol{u} - \boldsymbol{u}_t \cdot \boldsymbol{u}$

$$\frac{R}{\eta_{\rm eff}} \Delta \geq \boldsymbol{y}_k \cdot \boldsymbol{u} - (\boldsymbol{u}_t \cdot \boldsymbol{u})(\boldsymbol{y}_k \cdot \boldsymbol{u}_t) - \frac{\eta_{\rm eff}}{2R} \left(\|\boldsymbol{y}_k\|^2 \boldsymbol{u}_t \cdot \boldsymbol{u} + 2(\boldsymbol{y}_k \cdot \boldsymbol{u})(\boldsymbol{y}_k \cdot \boldsymbol{u}_t) \right)$$
$$- \frac{\eta_{\rm eff}^2}{2R^2} \|\boldsymbol{y}_k\|^2 \boldsymbol{y}_k \cdot \boldsymbol{u} \quad .$$

By employing (1), (5), (11) and (14) we get a lower bound on Δ

$$\Delta \geq \eta_{\rm eff} \left(\frac{\gamma_{\rm d}}{R} - \frac{\eta_{\rm eff}}{2} - \frac{\eta_{\rm eff}^2}{2} \right) - \eta_{\rm eff} \left(1 + \eta_{\rm eff}\right) \frac{\beta}{R} t^{-\epsilon} \quad . \tag{17}$$

From the misclassification condition it is obvious that convergence of the algorithm is impossible unless $C(t) \leq \gamma_{\rm d}$ i.e.

$$t \geq t_0 \equiv \left(\frac{\beta}{\gamma_{\rm d}}\right)^{\frac{1}{\epsilon}} \quad . \tag{18}$$

A repeated application of (17) $(t+1-t_0)$ times yields

$$\boldsymbol{u}_{t+1}\cdot\boldsymbol{u}-\boldsymbol{u}_{t_0}\cdot\boldsymbol{u} \geq \eta_{\text{eff}}\left(\frac{\gamma_{\text{d}}}{R}-\frac{\eta_{\text{eff}}}{2}-\frac{\eta_{\text{eff}}^2}{2}\right)(t+1-t_0)-\eta_{\text{eff}}\left(1+\eta_{\text{eff}}\right)\frac{\beta}{R}\sum_{m=t_0}^{t} m^{-\epsilon} \ .$$

By employing the inequality

$$\sum_{m=t_0}^{t} m^{-\epsilon} \leq \int_{t_0}^{t} m^{-\epsilon} dm + t_0^{-\epsilon} = \frac{t^{1-\epsilon}-t_0^{1-\epsilon}}{1-\epsilon} + t_0^{-\epsilon}$$

and taking into account (11) we finally obtain

$$1 \geq \eta_{\text{eff}}\left(\frac{\gamma_{\text{d}}}{R}-\frac{\eta_{\text{eff}}}{2}-\frac{\eta_{\text{eff}}^2}{2}\right)(t-t_0)-\eta_{\text{eff}}\left(1+\eta_{\text{eff}}\right)\frac{\beta}{R}\frac{\left(t^{1-\epsilon}-t_0^{1-\epsilon}\right)}{1-\epsilon} - \frac{\eta_{\text{eff}}^2}{2}\left(1+\eta_{\text{eff}}+2\frac{\gamma_{\text{d}}}{R}\right) \ . \quad (19)$$

Let us define the new variable $\tau \geq 0$ through the relation

$$t = t_0\left(1+\tau\right) = \left(\frac{\beta}{\gamma_{\text{d}}}\right)^{\frac{1}{\epsilon}}\left(1+\tau\right) \ . \quad (20)$$

In terms of τ (19) becomes

$$\frac{1}{\eta_{\text{eff}}}\left(\frac{\beta}{R}\right)^{-\frac{1}{\epsilon}}\left(\frac{\gamma_{\text{d}}}{R}\right)^{\left(\frac{1}{\epsilon}-1\right)}\left(1+\frac{\eta_{\text{eff}}^2}{2}\left(1+\eta_{\text{eff}}+2\frac{\gamma_{\text{d}}}{R}\right)\right) \geq \left(1-\frac{\eta_{\text{eff}}}{2}\left(1+\eta_{\text{eff}}\right)\frac{R}{\gamma_{\text{d}}}\right)\tau-\left(1+\eta_{\text{eff}}\right)\frac{(1+\tau)^{1-\epsilon}-1}{1-\epsilon} \ . \quad (21)$$

Let $g(\tau)$ be the r.h.s. of the above inequality. Since

$$\chi \equiv \left(\frac{\gamma_{\text{d}}}{R}-\frac{\eta_{\text{eff}}}{2}\left(1+\eta_{\text{eff}}\right)\right) > 0 \ ,$$

given that $\eta_{\text{eff}} < \frac{\gamma_{\text{d}}}{R} \leq 1$, it is not difficult to verify that $g(\tau)$ (with $\tau \geq 0$) is unbounded from above and has a single extremum, actually a minimum, at

$$\tau_{\min} = \left(1+\eta_{\text{eff}}\right)^{\frac{1}{\epsilon}}\left(1-\frac{\eta_{\text{eff}}}{2}\left(1+\eta_{\text{eff}}\right)\frac{R}{\gamma_{\text{d}}}\right)^{-\frac{1}{\epsilon}}-1 \geq 0$$

with $g(\tau_{\min}) \leq 0$. Moreover, the l.h.s of (21) is positive. Therefore, there is a single value τ_b of τ where (21) holds as an equality which provides an upper bound on τ

$$\tau \leq \tau_b \quad (22)$$

satisfying

$$\tau_b \geq \tau_{\min} \geq 0 \ .$$

Combining now (20) and (22) we obtain the bound on the number of updates

$$t \leq t_b \equiv \left(\frac{\beta}{\gamma_{\rm d}}\right)^{\frac{1}{\epsilon}} (1+\tau_b) \tag{23}$$

which provides a proof, alternative to the one along the lines of Sect. 2, that the algorithm converges in a finite number of steps. From (23) and taking into account the misclassification condition we obtain a lower bound β/t_b^{ϵ} on the margin achieved. Thus, the fraction f of the directional margin that the algorithm achieves satisfies

$$f \geq \frac{\beta/\gamma_{\rm d}}{{t_b}^{\epsilon}} = (1+\tau_b)^{-\epsilon} \quad . \tag{24}$$

Let us assume that $\frac{\beta}{R} \to \infty$ in which case from (15) $\eta_{\rm eff} \to 0$ and (21) becomes

$$\frac{1}{\eta_0}\left(\frac{\beta}{R}\right)^{-\left(\frac{1}{\epsilon}-\delta\right)}\left(\frac{\gamma_{\rm d}}{R}\right)^{\left(\frac{1}{\epsilon}-1\right)} \geq \tau - \frac{(1+\tau)^{1-\epsilon}-1}{1-\epsilon} \quad . \tag{25}$$

Provided $\epsilon\delta < 1$ the l.h.s. of the above inequality vanishes in the limit $\frac{\beta}{R} \to \infty$. Then, since $\tau_{\rm min}$ vanishes as well, the r.h.s. of the inequality becomes a strictly increasing function of τ and (25) obviously holds as an equality only for $\tau = \tau_b = 0$. Therefore,

$$\tau_b \to \tau_{\rm min} \to 0 \quad \text{as} \quad \frac{\beta}{\rm R} \to \infty \ . \tag{26}$$

Combining (24) with (26) and taking into account that $f \leq 1$ we conclude that

$$f \to 1 \quad \text{as} \quad \frac{\beta}{\rm R} \to \infty \ .$$

□

We now turn to special cases some of which are not covered by Thm. 1.

$\epsilon = \frac{1}{2}$: For this case we obtain an explicit upper bound on the number of updates by solving the quadratic equation derived from (19). Setting $\psi = \left(1 + \eta_{\rm eff} + 2\frac{\gamma_{\rm d}}{R}\right)$ we get

$$t \leq \left(\frac{\beta}{\gamma_{\rm d}}\right)^2 \left\{1 + \frac{\eta_{\rm eff}}{2}\frac{\psi}{\chi} + \sqrt{\left(\frac{\eta_{\rm eff}}{2}\frac{\psi}{\chi}\right)^2 + \frac{\gamma_{\rm d}^2}{\beta^2\eta_{\rm eff}}\frac{(2+\eta_{\rm eff}^2\psi)}{2\chi}}\right\}^2 \quad .$$

In the limit $\frac{\beta}{R} \to \infty$ the quantity in braces on the r.h.s. of the above inequality tends to unity provided $\eta_{\rm eff}$ scales with β according to (15) with $0 < \delta < 2$. This demonstrates explicitly the statement made in Thm. 1.

$\epsilon\delta = 1$: If $\epsilon\delta = 1$ the l.h.s. of (25) becomes $\frac{1}{\eta_0}\left(\frac{\gamma_{\rm d}}{R}\right)^{\left(\frac{1}{\epsilon}-1\right)}$ which does not vanish in the limit $\frac{\beta}{R} \to \infty$. Therefore, τ_b tends to a non-zero value depending on η_0. If, however, $\eta_0 \gg \left(\frac{\gamma_{\rm d}}{R}\right)^{\left(\frac{1}{\epsilon}-1\right)}$ the bound τ_b can become very small leading to a guaranteed fraction of the margin achieved very close to 1.

$\epsilon = \delta = 1$: In this case $\eta = \eta_0$ and as $\epsilon \to 1$ (25) becomes

$$\frac{1}{\eta} \geq \tau - \ln(1+\tau) \ .$$

For $\eta = 1$ we obtain the bound $\tau_b \simeq 2.15$ leading to a fraction of the maximum margin $f \geq (1+\tau_b)^{-1} \simeq 0.32$. By choosing larger values of the learning rate η we can make the value of the guaranteed fraction approach unity. In this particular case, however, it is possible to obtain better bounds on the number of updates leading to larger estimates for the guaranteed fraction of the margin by different proof techniques. Following the one introduced by Gentile [4] (see Appendix B) we can obtain for $\epsilon = 1$, provided the inequalities $\eta \left(1 + \eta^2 R^2/\beta^2\right)^{-1} \leq 1$ and $\eta < \beta\gamma_{\rm d}/\sqrt{6}R^2$ are satisfied, the upper bound

$$t \leq \frac{2}{\eta}\frac{\beta}{\gamma_{\rm d}}\left(1 + \frac{\eta\gamma_{\rm d}}{\beta}\right)\left(1 + \frac{\eta^2 R^2}{\beta^2}\right) + \frac{8}{3}\left(\frac{R}{\gamma_{\rm d}}\right)^2\left(1 + \frac{\eta\gamma_{\rm d}}{\beta}\right)^2\left(1 + \frac{\eta^2 R^2}{\beta^2}\right)^2 + 1 \tag{27}$$

on t and the lower bound

$$f \geq \frac{\eta}{2}\left\{\left(1 + \frac{\eta R}{\beta}\right)\left(1 + \frac{\eta^2 R^2}{\beta^2}\right) + \frac{4}{3}\frac{\eta R^2}{\beta\gamma_{\rm d}}\left(1 + \frac{\eta R}{\beta}\right)^2\left(1 + \frac{\eta^2 R^2}{\beta^2}\right)^2 + \frac{\eta R}{2\beta}\right\}^{-1} \tag{28}$$

on the fraction f of the margin achieved. In the limit $\frac{\beta}{R} \to \infty$ we see that $f \geq \frac{\eta}{2}$ which saturating the constraint on η could become $f \geq \frac{1}{2}$. By imposing the more relaxed constraint $\eta \left(1 + \eta^2 R^2/\beta^2\right)^{-1} \leq 2$ we can show that in the limit $\frac{\beta}{R} \to \infty$

$$f \geq \frac{2\eta}{3}\left(1 + \sqrt{1 + \frac{8}{3}\frac{\gamma_{\rm d}^2}{R^2}}\right)^{-1} \ . \tag{29}$$

In this limit the constraint on η allows η values as large as 2. This fact combined with the observation that the ratio $\gamma_{\rm d}/R$ can be made very small by placing the patterns at a larger distance from the origin in the augmented space leads to a guaranteed fraction $\frac{2}{3}$ of the margin for the largest allowed value of η. Thus, our earlier conclusion that for $\epsilon = \delta = 1$ the guaranteed fraction of the margin achieved as $\frac{\beta}{R} \to \infty$ increases with η is confirmed by this alternative technique.

4 Experiments

In this section we present the results of experiments performed in order to verify the theoretical statements made earlier and to evaluate the performance of the CRAMMA$^{\epsilon}$ algorithm in comparison with the other two well-known similar in spirit ones, namely the standard Perceptron algorithm with margin and the ALMA$_2$ algorithm (as modified in Appendix A). Our primary concern will be the ability of the algorithms to find the maximum margin. For the Perceptron algorithm the guaranteed fraction of the margin achieved in terms of $\frac{b}{\eta R^2}$ is

Table 1. Experimental results for the sonar data set. The directional margin γ'_{d}, the number of epochs (eps) and updates per epoch (up/ep) are given for the Perceptron, the ALMA$_2$ and the CRAMMA$^{\epsilon}$ ($\delta = 1$) algorithms.

Perceptron				ALMA$_2$				CRAMMA$^{\epsilon}$ $\epsilon = \frac{1}{2}, \eta = 0.001$			
$\frac{b}{\eta R^2}$	γ'_{d}	eps	up/ep	α	γ'_{d}	eps	up/ep	$\frac{\beta}{R}$	γ'_{d}	eps	up/ep
1	0.00578	18793	13.2	0.8	0.00512	21828	11.2	0.65	0.00550	19066	11.3
3	0.00709	42053	15.2	0.6	0.00694	62009	14.1	1.5	0.00707	50887	12.9
5.5	0.00746	73283	15.5	0.5	0.00740	109466	15.1	2	0.00746	79658	13.2
20	0.00780	252938	15.7	0.4	0.00777	207486	16.1	3	0.00782	153699	13.9
30	0.00785	376879	15.7	0.3	0.00800	438763	17.1	4	0.00800	247707	14.7
100	0.00791	1242783	15.8	0.2	0.00818	1154964	18.1	6.5	0.00819	582937	15.7
500	0.00793	6189654	15.8	0.1	0.00831	5350730	19.1	14	0.00832	2358082	17.5

Table 2. Experimental results for the sonar data set. The directional margin γ'_{d}, the number of epochs (eps) and updates per epoch (up/ep) are given for the CRAMMA$^{\epsilon}$ algorithm with $\epsilon = 1$ and the values $\eta = 1, 2, 5$ of the learning rate ($\delta = 1$).

$\frac{\beta}{R}$	$\eta = 1$			$\eta = 2$			$\eta = 5$		
	γ'_{d}	eps	up/ep	γ'_{d}	eps	up/ep	γ'_{d}	eps	up/ep
1000	0.00670	36815	15.6						
2000	0.00725	65415	16.1						
3000	0.00747	94909	16.2						
5000	0.00761	154057	16.3	0.00768	137493	18.1			
7000	0.00767	212595	16.4	0.00781	187984	18.2			
10000	0.00772	299476	16.5	0.00791	264053	18.3			
20000	0.00778	593527	16.5	0.00803	517399	18.4	0.00808	465477	20.3
40000	0.00782	1179629	16.5	0.00808	1024704	18.4	0.00822	905121	20.5
60000	0.00783	1766427	16.5	0.00810	1531253	18.4	0.00827	1346366	20.6

$\left(2 + \eta R^2/b\right)^{-1}$ [6, 7, 10] and tends to $\frac{1}{2}$ as $\frac{b}{\eta R^2} \to \infty$. For the ALMA$_2$ algorithm this fraction in terms of the parameter $\alpha \in (0, 1]$ is $1 - \alpha$ [4]. Everywhere the data are embedded in the augmented space at a distance $\rho = 1$ from the origin in the additional dimension.

First we analyse the training data set of the sonar classification problem as selected for the aspect-angle dependent experiment in [5]. It consists of 104 instances each with 60 attributes obtainable from the UCI repository. The results of our comparative study of the Perceptron, ALMA$_2$ and CRAMMA$^{\epsilon}$ ($\epsilon = \frac{1}{2}$) algorithms are presented in Table 1. We observe that although for values of the margin not too close to the maximum one the CRAMMA$^{\epsilon}$ algorithm is probably not the fastest, for values in the vicinity of the maximum margin it is certainly the fastest by far. Moreover, the data suggest that the Perceptron is not always able to obtain margins infinitely close to the maximum one. In Table 2 we present results obtained by the CRAMMA$^{\epsilon}$ algorithm with $\epsilon = 1$ and increasing but β-independent values of the learning rate η. The behaviour observed is perfectly consistent with the one suggested by the theoretical analysis. Finally, in Table

Table 3. Experimental results for the sonar data set. The directional margin $\gamma'_{\rm d}$, the number of epochs (eps) and updates per epoch (up/ep) are given for the CRAMMA$^{\epsilon}$ algorithm with $\epsilon = 2$ and a β-dependent learning rate $\eta = 0.4(\frac{\beta}{R})^{1-\delta}$ with $\delta = 0.3$.

$\frac{\beta}{R}$	10^6	10^7	10^8	10^9	10^{10}	10^{11}	10^{12}	10^{13}
$\gamma'_{\rm d}$	0.00103	0.00366	0.00552	0.00669	0.00737	0.00780	0.00810	0.00827
eps	8534	7243	14264	35849	98873	281397	821499	2443708
up/ep	8.3	14.7	18.5	21.1	23.0	24.9	26.4	27.8

3 we present results obtained by the CRAMMA$^{\epsilon}$ algorithm with $\epsilon = 2$ but a β-dependent η ($\delta = 0.3$). Notice that $\eta \to \infty$ but $\eta_{\rm eff} \to 0$ as $\frac{\beta}{R} \to \infty$.

We additionally analyse an artificial data set known as LS-10 with 1000 instances divided into two classes. Each instance has 10 attributes whose values are uniformly distributed in [0,1]. The attributes x_i of the instances belonging to the first class satisfy the inequality $x_1 + \ldots + x_5 < x_6 + \ldots + x_{10}$ with the attributes of the instances of the other satisfying the inverse inequality. In Table 4 we present the results of a comparative study of the Perceptron, ALMA$_2$ and CRAMMA$^{\epsilon}$ algorithms. It is apparent that the performance of the CRAMMA$^{\epsilon}$ algorithm on this data set is astonishingly good, beyond any expectation. Although analogous extraordinary results are obtainable for $\epsilon = \frac{1}{2}$ and small learning rates (which, however, do not enter the theoretically expected guaranteed fraction of the maximum margin in the limit $\frac{\beta}{R} \to \infty$) we decided to present the ones obtained for the parameter values $\epsilon = 1$ and $\eta \ll 1$ for which we theoretically anticipate only a tiny guaranteed fraction of the margin.

Table 4. Experimental results for the LS-10 data set. The directional margin $\gamma'_{\rm d}$, the number of epochs (eps) and updates per epoch (up/ep) are given for the Perceptron, the ALMA$_2$ and the CRAMMA$^{\epsilon}$ ($\delta = 1$) algorithms.

Perceptron				ALMA$_2$				CRAMMA$^{\epsilon}$ $\epsilon = 1, \eta = 0.02$			
$\frac{b}{\eta R^2}$	$\gamma'_{\rm d}$	eps	up/ep	α	$\gamma'_{\rm d}$	eps	up/ep	$\frac{\beta}{R}$	$\gamma'_{\rm d}$	eps	up/ep
1	0.00245	168994	4.1	0.9	0.00160	79792	3.5	75	0.00278	6119	12.2
3	0.00273	362957	5.2	0.7	0.00242	408972	3.8	100	0.00282	7591	13.0
10	0.00286	1032612	5.7	0.5	0.00270	1357880	4.8	400	0.00288	26305	14.5

5 Conclusions

We presented a new class of approximate large margin classifiers characterised by a constant effective learning rate. Our theoretical approach, having its roots in the concept of stepwise convergence, proved sufficiently powerful in establishing asymptotic convergence to the optimal hyperplane for a whole class of algorithms in which the misclassification condition is relaxed with an arbitrary power of the number of updates. Our analysis was also confirmed experimentally.

A The ALMA$_2$ Algorithm

For completeness we briefly review the ALMA$_2$ algorithm [4] slightly modified in order to accommodate patterns which are not normalised to unit length. The update rule is the one of (3) with $f_t = 1$ and $\eta_t = \eta/\sqrt{t}$. The length of the newly produced weight vector $\boldsymbol{a}_{t+1}$ is subsequently normalised to R only if it exceeds that value. The misclassification condition is given by (7) and the initial value of the weight vector is $\boldsymbol{a}_1 = \boldsymbol{0}$. Following [4] one can derive the relation

$$R \geq \|\boldsymbol{a}_{t+1}\| \geq \frac{\eta\gamma_{\mathrm{d}}}{A+1}\frac{t}{\sqrt{2A+t}} \ , \tag{30}$$

where $A = \eta\left(\eta/2 + b/R^2\right)$. From (30) one can easily show that $\|\boldsymbol{a}_t\|$ satisfies the inequalities $c_3\sqrt{t-1} \leq \|\boldsymbol{a}_t\| \leq R$, where $c_3 = \eta\gamma_{\mathrm{d}}\left((A+1)\sqrt{2A+1}\right)^{-1}$, to which we referred in Sect. 2. From (30) one also gets the bound

$$t \leq t_b \equiv \left(\frac{A+1}{\eta}\right)^2\left(\frac{R}{\gamma_{\mathrm{d}}}\right)^2 + 2(A+1) \ . \tag{31}$$

Using (7) and (31) one can show that the fraction of the directional margin achieved satisfies

$$f \geq \frac{1}{\gamma_{\mathrm{d}}}\frac{b}{R\sqrt{t_b}} \geq 1-\alpha \ ,$$

where $\alpha \in (0,1]$ is related to the parameters η and b as follows

$$\frac{b}{R^2} = \frac{1-\alpha}{\alpha}\left(\frac{1}{\eta} + \frac{3}{2}\eta\right) \ . \tag{32}$$

We can partially optimise the value of η by minimising the dominant term proportional to $(R/\gamma_{\mathrm{d}})^2$ on the r.h.s. of (31) keeping fixed either b or α. In the former case we obtain the value $\eta = \sqrt{2}$ (also employed in [4]) whereas in the latter we obtain the value

$$\eta = \sqrt{\frac{2}{3-2\alpha}} \ .$$

This is the value of η chosen in our experiments since it led to faster convergence and to larger margin values for fixed α. Once η is fixed b is determined from (32).

B Bounds for the CRAMMA$^{\epsilon}$ Algorithm with $\epsilon = \delta = 1$

In this appendix we sketch the derivation of (27), (28) and (29) following the technique of [4]. Taking the inner product of (3) with the optimal direction $\boldsymbol{u}$, employing (1) and repeatedly applying the resulting inequality we have

$$\begin{aligned}\beta = \|\boldsymbol{a}_{t+1}\| \geq \boldsymbol{a}_{t+1}\cdot\boldsymbol{u} &= \frac{\boldsymbol{a}_t\cdot\boldsymbol{u} + \eta\boldsymbol{y}_k\cdot\boldsymbol{u}}{N_{t+1}} \geq \frac{\boldsymbol{a}_t\cdot\boldsymbol{u}}{N_{t+1}} + \frac{\eta\gamma_{\mathrm{d}}}{N_{t+1}} \\ &\geq \frac{\boldsymbol{a}_1\cdot\boldsymbol{u}}{N_{t+1}N_t\cdots N_2} + \eta\gamma_{\mathrm{d}}\left(\frac{1}{N_{t+1}} + \frac{1}{N_{t+1}N_t} + \cdots + \frac{1}{N_{t+1}N_t\cdots N_2}\right) \ . \end{aligned} \tag{33}$$

For the normalisation factor N_{m+1} we can derive the inequality

$$N_{m+1}^{-1} \geq \alpha^{-1} \left(1 + 2A/m\right)^{-\frac{1}{2}} \equiv r_m \ ,$$

where $\alpha = \left(1 + \eta^2 R^2/\beta^2\right)^{\frac{1}{2}}$ and $A = \eta\alpha^{-2}$, which if substituted in (33) leads to

$$\frac{1}{\eta}\frac{\beta}{\gamma_{\rm d}} \geq \sum_{m=1}^{t}\prod_{j=m}^{t} r_j \geq \sum_{m=2}^{t}\prod_{j=m}^{t} r_j = r_t \sum_{m=2}^{t}\prod_{j=m}^{t-1} r_j \geq r_t \sum_{m=2}^{t} \alpha^{m-t}\left(\frac{m-1}{t-1}\right)^A \tag{34}$$

given that $\boldsymbol{a}_1 \cdot \boldsymbol{u} > 0$. At the last stage of the previous inequality we made use of

$$-\ln\prod_{j=m}^{t-1} r_j \leq (t-m)\ln\alpha + \sum_{j=m}^{t-1}\frac{A}{j} \leq \ln a^{t-m} + A\int_{m-1}^{t-1}\frac{dj}{j} \ .$$

Taking into account (18) and the fact that $A \leq \eta$ we have that $(1+2A/t)^{\frac{1}{2}} \leq (1+2\eta\gamma_{\rm d}/\beta)^{\frac{1}{2}} \leq 1 + \eta\gamma_{\rm d}/\beta$. Using the latter inequality and setting $l = m-1$ (34) gives

$$1 + \frac{1}{\eta}\frac{\beta}{\gamma_{\rm d}} \geq \alpha^{-t}(t-1)^{-A}\sum_{l=1}^{t-1} l^A \alpha^l \ . \tag{35}$$

Let us first assume that $A \leq 1$. Then, since $l/(t-1) \leq 1$, we can set $A = 1$ in (35) and using

$$\sum_{l=1}^{n} l\alpha^l = \alpha\frac{d}{d\alpha}\sum_{l=1}^{n}\alpha^l = \alpha\frac{d}{d\alpha}\left(\alpha\frac{\alpha^n - 1}{\alpha - 1}\right) = \frac{n\alpha^{n+1}}{(\alpha-1)^2}\left\{(\alpha-1) - \frac{1-\alpha^{-n}}{n}\right\}$$

obtain

$$\frac{1-\alpha^{-(t-1)}}{t-1} \geq (\alpha-1)\left\{1 - (\alpha-1)\left(1 + \frac{1}{\eta}\frac{\beta}{\gamma_{\rm d}}\right)\right\} \ . \tag{36}$$

The r.h.s. of (36) is certainly positive if $\eta_{\rm eff} < \gamma_{\rm d}/R$ or $\eta < \beta\gamma_{\rm d}/R^2$. Since the l.h.s. is a monotonically decreasing function of t vanishing in the limit $t \to \infty$ (36) gives rise to an upper bound on t. To obtain an approximation of this upper bound (i.e. obtain a less restrictive upper bound) we employ the relation $\alpha^{-(t-1)} = e^{-(t-1)\ln\alpha}$ and the inequalities $(1-e^{-x})/x \leq 1 - x/2 + x^2/6$ for $x > 0$, $(x-1) - (x-1)^2/2 \leq \ln x \leq (x-1)$ for $x > 1$ and $1/\ln x \leq x/(x-1)$ for $1 < x \leq 2$. Then, (36) can be shown to lead to

$$(t-1)^2 - \frac{3}{\alpha-1}(t-1) + 6\frac{\alpha^2}{\alpha-1}\left(1 + \frac{1}{\eta}\frac{\beta}{\gamma_{\rm d}}\right) \geq 0 \tag{37}$$

which gives the expected upper bound on $(t-1)$, namely the smallest positive root of the corresponding quadratic equation. Real roots exist if $\eta < \beta\gamma_{\rm d}/\sqrt{6}R^2$ and are approximated by using the inequality $\sqrt{1-x} \geq 1 - x/2 - x^2/2$. The bound obtained is the one of (27) from which (28) is readily derivable.

If $A \leq 2$, again because $l/(t-1) \leq 1$, we can set $A = 2$ in (35). Then, using

$$\sum_{l=1}^{n} l^2 \alpha^l = a \frac{d}{d\alpha} \sum_{l=1}^{n} l \alpha^l = \frac{n\alpha^{n+1}}{(\alpha-1)^3} \left\{ n(\alpha-1)^2 - 2(\alpha-1) + (\alpha+1) \frac{1-\alpha^{-n}}{n} \right\} \ ,$$

we get

$$\frac{\alpha-1}{t-1} - \frac{1-\alpha^{-(t-1)}}{(t-1)^2} \geq \frac{1}{2}(\alpha-1)^2 \left\{ 1 - (\alpha-1) \left(1 + \frac{1}{\eta} \frac{\beta}{\gamma_{\rm d}} \right) \right\} \ . \qquad (38)$$

The l.h.s of (38) can be shown to be a strictly decreasing function of t vanishing as $t \to \infty$ whereas its r.h.s is positive if $\eta < \beta\gamma_{\rm d}/R^2$. Thus, (38) leads to an upper bound on t. To find an approximation of such a bound we employ again the relation $\alpha^{-(t-1)} = e^{-(t-1)\ln\alpha}$ and the additional inequality $(1-e^{-x})/x \geq 1 - x/2 + x^2/6 - x^3/24$ for $x > 0$ in (38) to obtain the less restrictive relation

$$(t-1)^3 - \frac{4}{\alpha-1}(t-1)^2 + 12\frac{\alpha^2}{\alpha-1} \left(1 + \frac{1}{\eta} \frac{\beta}{\gamma_{\rm d}} \right) (t-1) + 12\frac{\alpha^4}{(\alpha-1)^2} \geq 0 \ .$$

In the limit $\frac{\beta}{R} \to \infty$ the above inequality is satisfied if t is bounded from above by the smallest positive root of the corresponding cubic equation. This leads to (29).

References

1. Aizerman, M. A., Braverman, E.M., Rozonoer, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control **25** (1964) 821–837
2. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines (2000) Cambridge, UK: Cambridge University Press
3. Duda, R.O., Hart, P.E.: Pattern Classsification and Scene Analysis (1973) Wiley-Interscience
4. Gentile C.:A new approximate maximal margin classification algorithm. Journal of Machine Learning Research **2** (2001) 213–242
5. Gorman, R. P., Sejnowski, T. J.: Analysis of hidden units in a layered network trained to classify sonar targets. Neural Networks **1** (1988) 75–89
6. Krauth, W., Mézard, M.: Learning algorithms with optimal stability in neural networks. Journal of Physics **A 20** (1987) L745–L752
7. Li, Y., Zaragoza, H., Herbrich, R., Shawe-Taylor, J., Kandola, J.: The perceptron algorithm with uneven margins. In ICML'02 379–386
8. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review **65**(6) (1958) 386–408
9. Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory **44**(5) (1998) 1926–1940
10. Tsampouka, P., Shawe-Taylor, J.: Analysis of generic perceptron-like large margin classifiers. ECML 2005, LNAI **3720** (2005) 750–758, Springer-Verlag
11. Vapnik, V. N.: The Nature of Statistical Learning Theory (1995) Springer Verlag