

The Margitron: A Generalised Perceptron with Margin

Constantinos Panagiotakopoulos and Petroula Tsampouka

Physics Division, School of Technology
Aristotle University of Thessaloniki, Greece
`costapan@eng.auth.gr, petroula@gen.auth.gr`

Abstract. We identify the classical Perceptron algorithm with margin as a member of a broader family of large margin classifiers which we collectively call the Margitron. The Margitron, (despite its) sharing the same update rule with the Perceptron, is shown in an incremental setting to converge in a finite number of updates to solutions possessing any desirable fraction of the maximum margin. Experiments comparing the Margitron with decomposition SVMs on tasks involving linear kernels and 2-norm soft margin are also reported.

1 Introduction

It is widely accepted that the larger the margin of the solution hyperplane the greater is the generalisation ability of the learning machine [18, 14]. The simplest online learning algorithm for binary linear classification, the Perceptron [12, 11], does not aim at any margin. The problem, instead, of finding the optimal margin hyperplane lies at the core of Support Vector Machines (SVMs) [18, 1]. Their efficient implementation, however, is somewhat hindered by the fact that they require solving a quadratic programming problem.

The complications encountered in implementing SVMs has resprurred the interest in alternative large margin classifiers many of which are based on the Perceptron algorithm. The oldest such algorithm which appeared long before the advent of SVMs is the standard Perceptron with margin [2], a straightforward extension of the Perceptron, which, however, in an incremental setting is known to be able to guarantee achieving only up to $1/2$ of the maximum margin that the dataset possesses [8, 10, 15]. Subsequently, various algorithms succeeded in achieving larger fractions of the maximum margin by employing modified perceptron-like update rules. Such algorithms include ROMMA [9], ALMA [3], CRAMMA [16] and MICRA [17]. A somewhat different approach from the hard margin one adopted by most of the algorithms above was also developed which focuses on the minimisation of the 1-norm soft margin loss through stochastic gradient descent. There is a connection, however, between such algorithms and the Perceptron since their unregularised form with constant learning rate is identical to the Perceptron with margin. Notable representatives of this approach are the pioneer NORMA [7] and the very recent Pegasos [13].

A question that arises naturally and which we attempt to answer in the present work is whether it is possible to achieve a guaranteed fraction of the maximum margin larger than $1/2$ while retaining the original perceptron update rule. To this end we construct a whole new family of algorithms at least one member of which has guaranteed convergence in a finite number of steps to a solution hyperplane possessing any desirable fraction of the unknown maximum margin. This family of algorithms in which the classical Perceptron with margin is naturally embedded will be termed the Margitron. Hopefully, the algorithms belonging to the margitron family by virtue of being generalisations of the very successful Perceptron will have a respectable performance in various classification tasks.

Section 2 contains some preliminaries and the description of the Margitron algorithm. Section 3 is devoted to a theoretical analysis. Section 4 contains our experimental results while Section 5 our conclusions.

2 The Margitron Algorithm

In what follows we assume that we are given a training set which either is linearly separable from the beginning or becomes separable by an appropriate feature mapping into a space of a higher dimension [18, 1]. This higher dimensional feature space in which the patterns are linearly separable will be the considered space. By placing all patterns in the same position at a distance ρ in an additional dimension we construct an embedding of our data into the so-called augmented space [2]. The advantage of this embedding is that the linear hypothesis in the augmented space becomes homogeneous. 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, $R \equiv \max_k \|\mathbf{y}_k\|$, with \mathbf{y}_k the k^{th} augmented pattern. Obviously, $R \geq \rho$.

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

$$\mathbf{u} \cdot \mathbf{y}_k \geq \gamma_d \equiv \max_{\mathbf{u}': \|\mathbf{u}'\|=1} \min_i \{\mathbf{u}' \cdot \mathbf{y}_i\} \quad \forall k. \quad (1)$$

We shall refer to γ_d as the maximum directional margin. It coincides with the maximum margin in the augmented space with respect to hyperplanes passing through the origin if no reflection is assumed. The directional margin γ_d and the maximum geometric margin γ in the original (non-augmented) feature space satisfy the inequality

$$1 \leq \gamma/\gamma_d \leq R/\rho.$$

As $\rho \rightarrow \infty$, $R/\rho \rightarrow 1$ and from the above inequality $\gamma_d \rightarrow \gamma$ [15].

In the Margitron algorithm the augmented weight vector \mathbf{a}_t is initially set to zero, i.e. $\mathbf{a}_0 = \mathbf{0}$, and is updated according to the classical perceptron rule

$$\mathbf{a}_{t+1} = \mathbf{a}_t + \mathbf{y}_k \quad (2)$$

each time a misclassification condition is satisfied by a training pattern \mathbf{y}_k . For the misclassification condition we consider two options. The first is to replace the constant functional margin threshold $b > 0$ in the misclassification condition of the classical Perceptron with margin by a term proportional to a power of the number of steps (updates) t

$$\mathbf{a}_t \cdot \mathbf{y}_k \leq b t^{1-\epsilon} , \quad \epsilon > 0 . \quad (3)$$

As a second option we employ a margin threshold proportional to a power of the length of the augmented weight vector leading to a misclassification condition

$$\mathbf{a}_t \cdot \mathbf{y}_k \leq b \|\mathbf{a}_t\|^{1-\epsilon} , \quad \epsilon > 0 . \quad (4)$$

For $t = 0$ in both (3) and (4) the threshold is set to 0 resulting in the first pattern being always misclassified. The Margitron with misclassification condition given by (3) will be referred to as the t -margitron whereas the version with condition given by (4) as the ℓ -margitron. Setting $\epsilon = 1$ in both the t - and the ℓ -margitron we recover the Perceptron with margin. Notice that the introduction of a constant learning rate is pointless since it amounts to a rescaling of b .

t-margitron	ℓ -margitron
Input: A linearly separable augmented set $S = (\mathbf{y}_1, \dots, \mathbf{y}_k, \dots, \mathbf{y}_m)$ with reflection assumed Fix: ϵ, b Define: $\bar{\epsilon} = 1 - \epsilon$ Initialise: $t = 0, \mathbf{a}_0 = \mathbf{0}, b_0 = 0$ repeat for $k = 1$ to m do $p_{tk} = \mathbf{a}_t \cdot \mathbf{y}_k$ if $p_{tk} \leq b_t$ then $\mathbf{a}_{t+1} = \mathbf{a}_t + \mathbf{y}_k$ $t \leftarrow t + 1$ $b_t = b t^{\bar{\epsilon}}$ end if end for until no update made within the for loop	Input: A linearly separable augmented set $S = (\mathbf{y}_1, \dots, \mathbf{y}_k, \dots, \mathbf{y}_m)$ with reflection assumed Fix: ϵ, b Define: $q_k = \ \mathbf{y}_k\ ^2, \bar{\epsilon} = \frac{1}{2}(1 - \epsilon)$ Initialise: $t = 0, \mathbf{a}_0 = \mathbf{0}, \ell_0 = 0, b_0 = 0$ repeat for $k = 1$ to m do $p_{tk} = \mathbf{a}_t \cdot \mathbf{y}_k$ if $p_{tk} \leq b_t$ then $\mathbf{a}_{t+1} = \mathbf{a}_t + \mathbf{y}_k$ $\ell_{t+1} = \ell_t + 2p_{tk} + q_k$ $t \leftarrow t + 1$ $b_t = b \ell_t^{\bar{\epsilon}}$ end if end for until no update made within the for loop

Fig. 1. The algorithms t -margitron and ℓ -margitron.

Both (3) and (4) can be written for $t > 0$ in the form

$$\mathbf{u}_t \cdot \mathbf{y}_k \leq C(t) \quad (5)$$

($\mathbf{u}_t \equiv \mathbf{a}_t / \|\mathbf{a}_t\|, C(t) > 0$) involving the margin $\mathbf{u}_t \cdot \mathbf{y}_k$ in the augmented space of the pattern \mathbf{y}_k with respect to the zero-threshold hyperplane normal to \mathbf{a}_t (i.e. the directional margin of \mathbf{y}_k) instead of its functional margin $\mathbf{a}_t \cdot \mathbf{y}_k$. The function

$C(t)$ is given by $C(t) = bt^{1-\epsilon} \|\mathbf{a}_t\|^{-1}$ for the t -margintron and by $C(t) = b \|\mathbf{a}_t\|^{-\epsilon}$ for the ℓ -margintron. We expect that $\epsilon < 1$ will result in an enhancement of the margin threshold $C(t)$ relative to the case $\epsilon = 1$ (Perceptron with margin) and that this enhancement will eventually lead to a slower average fall off of $C(t)$ with t progressing instead of a genuine increase which is desirable in order for the algorithm to converge. This expectation is further supported by the fact that, as we demonstrate below, $C(t) \leq ct^{-\epsilon}$ with $c > 0$. Hopefully, such a slower decrease of the margin required by the misclassification condition will ensure convergence to solutions possessing margins which are larger fractions of γ_d .

Taking the inner product of (2) with the optimal direction \mathbf{u} we obtain

$$\mathbf{a}_{t+1} \cdot \mathbf{u} - \mathbf{a}_t \cdot \mathbf{u} = \mathbf{y}_k \cdot \mathbf{u} \geq \gamma_d$$

a repeated application of which gives [11]

$$\|\mathbf{a}_t\| \geq \mathbf{a}_t \cdot \mathbf{u} \geq \gamma_d t. \quad (6)$$

Using (6) we get $C(t) \leq ct^{-\epsilon}$ with $c = b\gamma_d^{-1}$ and $c = b\gamma_d^{-\epsilon}$ for the t - and the ℓ -margintron, respectively.

3 Theoretical Analysis

Lemma 1. *Let*

$$g(t) = t^\epsilon - \alpha t^{\epsilon-1} - \beta$$

with $t \in [1, +\infty)$, $\epsilon > 0$, $\alpha \geq 1$ and $\beta > 0$. Then, there is a single value t_b of t satisfying

$$g(t_b) = 0$$

which is bounded as follows

$$\begin{aligned} \alpha + \beta^{\frac{1}{\epsilon}} \leq t_b \leq \frac{1}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}} & \quad \epsilon \leq 1 \\ \frac{1}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}} < t_b < \alpha + \beta^{\frac{1}{\epsilon}} & \quad \epsilon > 1. \end{aligned}$$

Proof. The function $g(t)$ with $g(1) < 0$ is unbounded from above and is either strictly increasing (if $\epsilon \leq 1$) or has at most one local minimum (if $\epsilon > 1$). Therefore, there is a single root t_b of $g(t)$. In addition, for $g(t) \neq 0$ $\text{sign}(t - t_b) = \text{sign}(g(t))$. Let $0 < \epsilon < 1$. We have $g(\alpha + \beta^{\frac{1}{\epsilon}}) = \beta^{\frac{1}{\epsilon}}(\alpha + \beta^{\frac{1}{\epsilon}})^{\epsilon-1} - \beta < \beta^{\frac{1}{\epsilon}}\beta^{\frac{\epsilon-1}{\epsilon}} - \beta = 0$, implying that $t_b > \alpha + \beta^{\frac{1}{\epsilon}}$. Moreover, $g(\frac{1}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}}) = (\frac{1-\epsilon}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}})(\frac{1}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}})^{\epsilon-1} - \beta = \beta(1 + \frac{1-\epsilon}{\epsilon} \alpha \beta^{-\frac{1}{\epsilon}})(1 + \frac{1}{\epsilon} \alpha \beta^{-\frac{1}{\epsilon}})^{\epsilon-1} - \beta > \beta(1 + \frac{1}{\epsilon} \alpha \beta^{-\frac{1}{\epsilon}})^{1-\epsilon}(1 + \frac{1}{\epsilon} \alpha \beta^{-\frac{1}{\epsilon}})^{\epsilon-1} - \beta = 0$ implying that $t_b < \frac{1}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}}$. (Here we make use of $1 + qz > (1 + z)^q$ for $-1 < z \neq 0$ and $0 < q < 1$.) If $\epsilon > 1$, instead, both the above inequalities are reversed. (Here we make use of $1 - qz < (1 + z)^{-q}$ for $z, q > 0$.) Finally, for $\epsilon = 1$ obviously $t_b = \alpha + \beta$. \square

Theorem 1. *The t -margitron with $0 < \epsilon \leq 1$ converges in*

$$t_c \leq \frac{1}{\epsilon} \frac{R^2}{\gamma_d^2} + \left(\frac{2}{2-\epsilon} \frac{b}{\gamma_d^2} \right)^{\frac{1}{\epsilon}} \quad (7)$$

updates to a solution hyperplane possessing directional margin γ'_d which is a fraction f of the maximum directional margin γ_d obeying the inequality

$$f \equiv \frac{\gamma'_d}{\gamma_d} \geq \left(\frac{R^2}{b} + \frac{2}{2-\epsilon} \right)^{-1} . \quad (8)$$

Moreover, an after-running estimate of $\frac{\gamma'_d}{\gamma_d}$ is obtainable from

$$\frac{\gamma'_d}{\gamma_d} \geq f_{\text{est}} \equiv \left(\frac{R^2}{b} t_c^{\epsilon-1} + \frac{2}{2-\epsilon} \right)^{-1} . \quad (9)$$

Proof. From (2) and taking into account (3) we get

$$\|\mathbf{a}_{t+1}\|^2 - \|\mathbf{a}_t\|^2 = \|\mathbf{y}_k\|^2 + 2\mathbf{y}_k \cdot \mathbf{a}_t \leq R^2 + 2bt^{1-\epsilon}$$

a repeated application t times of which leads to

$$\|\mathbf{a}_t\|^2 \leq R^2 t + 2b \sum_{l=1}^{t-1} l^{1-\epsilon} \leq R^2 t + 2b \int_0^t l^{1-\epsilon} dl = R^2 t + \frac{2}{2-\epsilon} b t^{2-\epsilon} . \quad (10)$$

Combining (6) with (10) we obtain

$$\gamma_d t \leq \|\mathbf{a}_t\| \leq R \sqrt{t + \frac{2}{2-\epsilon} \frac{b}{R^2} t^{2-\epsilon}} \quad (11)$$

from where

$$t^\epsilon \leq \frac{R^2}{\gamma_d^2} t^{\epsilon-1} + \frac{2}{2-\epsilon} \frac{b}{\gamma_d^2} \quad (12)$$

or, equivalently,

$$g(t) \equiv t^\epsilon - \frac{R^2}{\gamma_d^2} t^{\epsilon-1} - \frac{2}{2-\epsilon} \frac{b}{\gamma_d^2} \leq 0 . \quad (13)$$

The value t_b of t for which the above relation holds as an equality provides an upper bound on the number of updates t_c required for convergence. According to Lemma 1 there is a single such value which is bounded as stated there. This leads to the looser bound of (7).

Combining (3) with (5) and using (10) we obtain

$$\frac{C(t)}{\gamma_d} = \frac{b t^{1-\epsilon}}{\gamma_d \|\mathbf{a}_t\|} \geq \left(\frac{\gamma_d R}{b} \sqrt{t^{2\epsilon-1} + \frac{2}{2-\epsilon} \frac{b}{R^2} t^\epsilon} \right)^{-1} . \quad (14)$$

Multiplying both sides of (12) with its r.h.s. we get

$$t^\epsilon \left(\frac{R^2}{\gamma_d^2} t^{\epsilon-1} + \frac{2}{2-\epsilon} \frac{b}{\gamma_d^2} \right) \leq \left(\frac{R^2}{\gamma_d^2} t^{\epsilon-1} + \frac{2}{2-\epsilon} \frac{b}{\gamma_d^2} \right)^2 ,$$

or

$$\frac{\gamma_d R}{b} \sqrt{t^{2\epsilon-1} + \frac{2}{2-\epsilon} \frac{b}{R^2} t^\epsilon} \leq \frac{R^2}{b} t^{\epsilon-1} + \frac{2}{2-\epsilon}.$$

Using this last inequality and taking into account that $f = \gamma'_d/\gamma_d \geq C(t_c)/\gamma_d$ (14) leads to (9). Setting $t_c = 1$ in (9) we obtain the weaker bound of (8). \square

Remark 1. Noticing that the number of updates t_c required for convergence of the t-margitron satisfies (12) we get

$$\gamma_d \leq R \sqrt{t_c^{-1} + \frac{2}{2-\epsilon} \frac{b}{R^2} t_c^{-\epsilon}}$$

from where an alternative after-running lower bound on γ'_d/γ_d is obtainable. This bound, however, does not have to be smaller than $1 - \frac{\epsilon}{2}$.

Remark 2. The r.h.s. of (14) has in the interval $[1, +\infty)$ a single extremum, which is a maximum, at $t_\star = \left(|1 - 2\epsilon|(2-\epsilon)(2\epsilon)^{-1} \frac{R^2}{b} \right)^{\frac{1}{1-\epsilon}} \text{sign}(1 - 2\epsilon)$. Therefore, it is legitimate in calculating a lower bound on $C(t_c)/\gamma_d$ using (14) to replace t_c with t_b provided $t_c \geq t_\star$. This leads to the stronger than the one of (8) bound

$$f \geq \left(\frac{R^2}{b} t_b^{\epsilon-1} + \frac{2}{2-\epsilon} \right)^{-1} \quad (15)$$

which, however, is γ_d -dependent. The condition $t_c \geq t_\star$ is automatically satisfied for $\frac{1}{2} \leq \epsilon \leq 1$. For $0 < \epsilon < \frac{1}{2}$, instead, we may ensure that $t_c \geq t_\star$ if the r.h.s. of (14) is larger than or equal to 1 for $t = 1$ and as a consequence the normalised margin threshold $C(t)$ is initially not lower than the maximum directional margin γ_d , i.e. $C(1) \geq \gamma_d$. A condition sufficient for this to be the case is $\frac{b}{R^2} \geq \frac{\gamma_d}{R} \left(1 + \frac{2}{2-\epsilon} \frac{\gamma_d}{R} \right)$. In this event the algorithm is forced to converge only after $C(t)$ has fallen below γ_d which cannot occur as long as $t < t_\star$. If we choose

$$\frac{b}{R^2} = \left(1 - \frac{\epsilon}{2} \right)^{1-\epsilon} \delta^{-\epsilon} \left(\frac{\gamma_d^2}{R^2} \right)^{1-\epsilon} \quad (16)$$

and replace in (15) t_b with its lower bound $t_{lb} \equiv \left(\frac{2}{2-\epsilon} \frac{b}{\gamma_d^2} \right)^{\frac{1}{\epsilon}} = \frac{2}{2-\epsilon} \delta^{-1} \frac{R^2}{\gamma_d^2}$, which is lower than the lower bound inferred from Lemma 1, we can easily verify that $f \geq \left(\delta + \frac{2}{2-\epsilon} \right)^{-1}$. If $0 < \epsilon < \frac{1}{2}$ the parameter δ should satisfy the constraint $\delta \leq \left(1 - \frac{\epsilon}{2} \right)^{\frac{1-\epsilon}{\epsilon}} \left(\frac{\gamma_d}{R} \right)^{\frac{1}{\epsilon}-2} \left(1 + \frac{2}{2-\epsilon} \frac{\gamma_d}{R} \right)^{-\frac{1}{\epsilon}}$ which for $0 < \epsilon \ll \frac{1}{2}$ and $\frac{\gamma_d}{R} \ll 1$ suggests a rather slow convergence. Thus, it is not advisable in this case to employ values of b for which the constraint on δ is satisfied. The algorithm will still be able to achieve a large fraction of γ_d if it happens to converge in a sufficiently large number of updates t_c as it can be deduced from (9).

Lemma 2. For $x, y > 0$ and $-1 < \epsilon \leq 1$ it holds that

$$\frac{x^{1+\epsilon}}{1+\epsilon} - \frac{y^{1+\epsilon}}{1+\epsilon} \leq \frac{x^2 - y^2}{2y^{1-\epsilon}}. \quad (17)$$

Proof. For $\epsilon = 1$ (17) holds obviously as an equality. For $-1 < \epsilon < 1$ (17) is equivalent to $1 \leq \frac{1+\epsilon}{2}\alpha^{1-\epsilon} + \frac{1-\epsilon}{2}\alpha^{-(1+\epsilon)}$, with $\alpha = x/y$. The r.h.s. of the above inequality is minimised for $\alpha = 1$ and takes the value 1. \square

Lemma 3. For $t \geq 1$ and $0 < \epsilon \leq 1$ it holds that

$$\frac{t^\epsilon - 1}{\epsilon} \leq t^\epsilon (\ln t)^{1-\epsilon} - [\epsilon], \quad (18)$$

where $[\epsilon]$ denotes the integer part of ϵ .

Proof. For $t = 1$ or $\epsilon = 1$ (18) holds obviously as an equality. Let $t > 1$ and $0 < \epsilon < 1$. Then, with $x = t^\epsilon$ (18), as a strict inequality, is equivalent to $f(\epsilon) = \epsilon^\epsilon x (\ln x)^{1-\epsilon} - x + 1 > 0$. For $x \geq e^e$ we have $\frac{df}{d\epsilon} < 0$ from where $f(\epsilon) > \lim_{\epsilon \rightarrow 1} f(\epsilon) = 1$. For $1 < x < e^e$, instead, f has only one local minimum at $\epsilon = e^{-1} \ln x$ with value at that minimum given by $h(x) = x^{\frac{e-1}{e}} \ln x - x + 1$. It can be easily shown that $\frac{dh}{dx} = (1-e)x^{-e-1} \ln x^{-e-1} + x^{-e-1} - 1$ has no local minima in the interval $(1, e^e)$. Thus, $\frac{dh}{dx} > \min \left\{ \lim_{x \rightarrow 1} \frac{dh}{dx}, \lim_{x \rightarrow e^e} \frac{dh}{dx} \right\} = 0$. Therefore, $h(x) > \lim_{x \rightarrow 1} h(x) = 0$ and consequently $f(\epsilon) > 0$. \square

Lemma 4. Let

$$g(t) = t^\epsilon - \left(\alpha_1 \left(\frac{\ln t}{t} \right)^{1-\epsilon} + \alpha_2 t^{-1} \right) - \beta$$

with $t \in [1, +\infty)$, $0 < \epsilon < 1$, $\alpha_1, \alpha_2, \beta > 0$ and $\alpha \equiv \alpha_1 + \alpha_2 \geq 2 + \epsilon$. Then, $g(t_0) > 0$ with

$$t_0 = \left(\frac{1}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}} \right) \left(\ln \left(\frac{1}{\epsilon} \alpha + \beta^{\frac{1}{\epsilon}} \right) \right)^{1-\epsilon}.$$

Proof. Let $\lambda = \alpha / (\alpha + \epsilon \beta^{\frac{1}{\epsilon}}) < 1$ and $x = \ln \frac{\alpha}{\lambda \epsilon} \geq \zeta \equiv \ln \left(1 + \frac{2}{\epsilon} \right) > 1$ such that $t_0 = \frac{\alpha}{\lambda \epsilon} x^{1-\epsilon} \geq 1 + \frac{2}{\epsilon} > e$, $\ln t_0 = x + (1-\epsilon) \ln x > 1$, $\beta = \alpha^\epsilon (1-\lambda)^\epsilon / (\lambda \epsilon)^\epsilon$ and $\alpha_2 t_0^{-1} < \alpha_2 (\ln t_0 / t_0)^{1-\epsilon}$. Then,

$$\begin{aligned} g(t_0) &> t_0^\epsilon - \alpha \left(\frac{\ln t_0}{t_0} \right)^{1-\epsilon} - \beta = t_0^\epsilon \left(1 - \lambda \epsilon \left(1 + (1-\epsilon) \frac{\ln x}{x} \right)^{1-\epsilon} - \frac{(1-\lambda)^\epsilon}{x^{(1-\epsilon)\epsilon}} \right) \\ &> 1 - \lambda \epsilon (1 + (1-\epsilon)e^{-1})^{1-\epsilon} - (1-\lambda)^\epsilon \zeta^{(\epsilon-1)\epsilon}. \end{aligned}$$

Here we made use of $t_0^\epsilon > 1$, $\frac{\ln x}{x} \leq e^{-1}$ and $\frac{1}{x^{(1-\epsilon)\epsilon}} \leq \zeta^{(\epsilon-1)\epsilon}$. This last expression is minimised with respect to λ for $\lambda = \frac{1 + (1-\epsilon)e^{-1} - \zeta^{-\epsilon}}{1 + (1-\epsilon)e^{-1}}$ which substituted leads to

$$g(t_0) > (1 + (1-\epsilon)e^{-1})^{-\epsilon} f(\epsilon)$$

with $f(\epsilon) \equiv (1-\epsilon)(1-\zeta^{-\epsilon}) + (1 + (1-\epsilon)e^{-1})^\epsilon - (1 + \epsilon(1-\epsilon)e^{-1})$. Employing the expansion $\ln z = \sum_{k=1}^{\infty} \frac{2}{2k-1} \left(\frac{z-1}{z+1}\right)^{2k-1}$ for $z > 0$ [4] we obtain $\zeta > \left(\frac{1+\epsilon}{2}\right)^{-1}$ from where $\zeta^{-\epsilon} < \left(\frac{1+\epsilon}{2}\right)^\epsilon = \left(1 - \frac{1}{2}(1-\epsilon)\right)^\epsilon < 1 - \frac{1}{2}\epsilon(1-\epsilon)$. Moreover, $(1 + (1-\epsilon)e^{-1})^\epsilon - (1 + \epsilon(1-\epsilon)e^{-1}) > -\frac{1}{2}\epsilon(1-\epsilon)^3 e^{-2}$ since $(1+z)^q - (1+qz) > \frac{1}{2}q(q-1)z^2$ for $z > 0$ and $0 < q < 1$. Thus, $f(\epsilon) > \frac{1}{2}\epsilon(1-\epsilon)^2(1 - (1-\epsilon)e^{-2}) > 0$ leading to $g(t_0) > 0$. \square

Theorem 2. *The ℓ -margintron with $0 < \epsilon \leq 1$ converges in*

$$t_c \leq \left(\frac{1}{\epsilon}A + B\frac{1}{\epsilon}\right) \left(\ln\left(\frac{1}{\epsilon}A + B\frac{1}{\epsilon}\right)\right)^{1-\epsilon} \quad (19)$$

updates, with $A = (2 + \epsilon - 2[\epsilon])\frac{R^2}{\gamma_d^2}$ and $B = (1 + \epsilon)\frac{b}{\gamma_d^{1+\epsilon}}$, to a solution hyperplane possessing directional margin γ'_d which is a fraction f of the maximum directional margin γ_d obeying the inequality

$$f \equiv \frac{\gamma'_d}{\gamma_d} \geq \left\{ \frac{(1+\epsilon)^{2\epsilon-[\epsilon]}}{(2\epsilon)^\epsilon} \left(\frac{R^{1+\epsilon}}{b}\right) + 1 + \epsilon \right\}^{-1}. \quad (20)$$

Moreover, for $0 < \epsilon < 1$ an after-running estimate of $\frac{\gamma'_d}{\gamma_d}$ is obtainable from

$$\frac{\gamma'_d}{\gamma_d} \geq f_{\text{est}} \equiv \left\{ \frac{R^{1+\epsilon}}{b} \left(N^{1+\epsilon} + \frac{1+\epsilon}{2\epsilon} \left(\frac{R}{\gamma'_d}\right)^{1-\epsilon} \left(t_c^\epsilon - \frac{N-\epsilon}{N^{1-\epsilon}}\right) t_c^{-1} + 1 + \epsilon \right\}^{-1}. \quad (21)$$

Here the integer $N > 0$ satisfies any of the constraints

$$t_c \geq N \geq \frac{1+\epsilon}{2} \left(\frac{R}{\gamma'_d}\right)^{1-\epsilon}, \quad t_c \geq N \left(\frac{1-\epsilon N^{-1}}{1-\epsilon}\right)^{\frac{1}{\epsilon}}. \quad (22)$$

Obviously, the choice $N = 1$ is always acceptable. A near optimal choice of N is $N_{\text{opt}} = \left\lceil \frac{1}{2} \left(\frac{R}{\gamma'_d}\right)^{1-\epsilon} \right\rceil + 1$, provided it satisfies one of the above constraints.

Proof. From (2) and taking into account (4) we get

$$\|\mathbf{a}_{t+1}\|^2 - \|\mathbf{a}_t\|^2 = \|\mathbf{y}_k\|^2 + 2\mathbf{y}_k \cdot \mathbf{a}_t \leq R^2 + 2b \|\mathbf{a}_t\|^{1-\epsilon}$$

or, assuming $t \geq 1$,

$$\frac{\|\mathbf{a}_{t+1}\|^2 - \|\mathbf{a}_t\|^2}{2\|\mathbf{a}_t\|^{1-\epsilon}} \leq \frac{R^2}{2\|\mathbf{a}_t\|^{1-\epsilon}} + b.$$

By using (6) in the r.h.s. of the above inequality and (17) in its l.h.s. we obtain

$$\frac{\|\mathbf{a}_{t+1}\|^{1+\epsilon}}{1+\epsilon} - \frac{\|\mathbf{a}_t\|^{1+\epsilon}}{1+\epsilon} \leq \frac{1}{2} \frac{R^2}{\gamma_d^{1-\epsilon}} t^{\epsilon-1} + b$$

a repeated application $t - N$ times ($t > N \geq 1$) of which gives

$$\begin{aligned} \frac{\|\mathbf{a}_t\|^{1+\epsilon}}{1+\epsilon} - \frac{\|\mathbf{a}_N\|^{1+\epsilon}}{1+\epsilon} &\leq \frac{1}{2} \frac{R^2}{\gamma_d^{1-\epsilon}} \sum_{l=N}^{t-1} l^{\epsilon-1} + b(t-N) \\ &\leq \frac{1}{2} \frac{R^2}{\gamma_d^{1-\epsilon}} \left(N^{\epsilon-1} + \int_{l=N}^{t-1} l^{\epsilon-1} dl \right) + bt \\ &= \frac{1}{2} \frac{R^2}{\gamma_d^{1-\epsilon}} \left(N^{\epsilon-1} + \frac{1}{\epsilon} (t-1)^\epsilon - \frac{1}{\epsilon} N^\epsilon \right) + bt \\ &\leq \frac{1}{2} \frac{R^2}{\gamma_d^{1-\epsilon}} \left(N^{\epsilon-1} + \frac{1}{\epsilon} t^\epsilon - [\epsilon] - \frac{1}{\epsilon} N^\epsilon \right) + bt . \end{aligned}$$

Thus, employing the obvious bound $\|\mathbf{a}_N\| \leq RN$, we are led to

$$\|\mathbf{a}_t\| \leq L_{Nt} \equiv R \left(N^{1+\epsilon} + \frac{1+\epsilon}{2\epsilon} \left(\frac{R}{\gamma_d} \right)^{1-\epsilon} \left(t^\epsilon - \frac{N - \epsilon + [\epsilon]}{N^{1-\epsilon}} \right) + (1+\epsilon) \frac{b}{R^{1+\epsilon}} t \right)^{\frac{1}{1+\epsilon}} \quad (23)$$

which, although derived for $t > N$, turns out to be satisfied even for $t = N$. Combining (6) with (23) we obtain

$$\gamma_d^{1+\epsilon} t^{1+\epsilon} \leq \|\mathbf{a}_t\|^{1+\epsilon} \leq (L_{Nt})^{1+\epsilon} \quad (24)$$

from where

$$t^\epsilon \leq \left(\frac{L_{Nt}}{\gamma_d} \right)^{1+\epsilon} t^{-1} \quad (25)$$

or, equivalently,

$$g_N(t) \leq 0 \quad (26)$$

with

$$\begin{aligned} g_N(t) &\equiv t^\epsilon - \left(\frac{L_{Nt}}{\gamma_d} \right)^{1+\epsilon} t^{-1} \\ &= t^\epsilon - \left(\frac{R}{\gamma_d} \right)^{1+\epsilon} N^{1+\epsilon} t^{-1} - \frac{1+\epsilon}{2\epsilon} \frac{R^2}{\gamma_d^2} \left(t^\epsilon - N^\epsilon + (\epsilon - [\epsilon]) N^{\epsilon-1} \right) t^{-1} - (1+\epsilon) \frac{b}{\gamma_d^{1+\epsilon}} . \end{aligned} \quad (27)$$

Let us consider the derivative of $g_N(t)$

$$\frac{dg_N}{dt} = D_N(t) t^{-2} ,$$

where

$$D_N(t) = \epsilon t^{1+\epsilon} + \left(\frac{R}{\gamma_d} \right)^{1+\epsilon} N^{1+\epsilon} + \frac{1+\epsilon}{2\epsilon} \frac{R^2}{\gamma_d^2} \left((1-\epsilon) t^\epsilon - N^\epsilon + (\epsilon - [\epsilon]) N^{\epsilon-1} \right) .$$

$D_N(t)$ is strictly increasing and therefore has at most one root t_{r_N} ($D_N(t_{r_N}) = 0$) where obviously $g_N(t)$ acquires a minimum (since $g_N(t)$ is unbounded from above) with $g_N(t_{r_N}) < 0$ (since $g_N(N) < 0$). Thus, $g_N(t)$ starts from negative values at $t = N$ and with t increasing either tends monotonically to infinity or decreases further until it acquires a minimum at $t = t_{r_N}$ and then increases monotonically towards infinity. In both cases there is a single value t_{b_N} of t for which

$$g_N(t_{b_N}) = 0 \quad (28)$$

and moreover for $g_N(t) \neq 0$

$$\text{sign}(t - t_{b_N}) = \text{sign}(g_N(t)) . \quad (29)$$

The unique value t_{b_N} of t for which (24), (25) and (26) hold as equalities provides an upper bound on the number of updates t_c required for convergence.

Combining (4), (5), (23), (27) and (28) we get

$$\begin{aligned} f &= \frac{\gamma'_d}{\gamma_d} \geq \frac{C(t_c)}{\gamma_d} = \frac{b}{\gamma_d \|a_{t_c}\|^\epsilon} \geq \frac{b}{\gamma_d L_{N t_c}^\epsilon} \geq \frac{b}{\gamma_d L_{N t_{b_N}}^\epsilon} = \frac{b}{\gamma_d^{1+\epsilon} t_{b_N}^\epsilon} \\ &= \left\{ \frac{R^{1+\epsilon}}{b} \left(N^{1+\epsilon} + \frac{1+\epsilon}{2\epsilon} \left(\frac{R}{\gamma_d} \right)^{1-\epsilon} \left(t_{b_N}^\epsilon - \frac{N - \epsilon + [\epsilon]}{N^{1-\epsilon}} \right) \right) t_{b_N}^{-1} + 1 + \epsilon \right\}^{-1} . \end{aligned} \quad (30)$$

For $\epsilon = 1$ the above lower bound on f is optimised for $N = 1$ in which case it reduces to (20). For $0 < \epsilon < 1$ we may replace in the above lower bound on f first γ_d with γ'_d and subsequently, on the condition that one of the constraints (22) is satisfied, t_{b_N} with t_c since both replacements can be shown to loosen the bound. Thus, we obtain $f \geq f_{\text{est}}$ with f_{est} given by (21). An approximate maximisation of f_{est} with respect to N leads to the near optimal value N_{opt} of Theorem 2.

Let us choose $N = 1$ in (27) and replace $\left(\frac{R}{\gamma_d}\right)^{1+\epsilon}$ with $\frac{R^2}{\gamma_d^2}$, thereby lowering the value of $g_1(t)$

$$g_1(t) \geq t^\epsilon - \frac{1+\epsilon}{2} \frac{R^2}{\gamma_d^2} \left(\frac{t^\epsilon - 1}{\epsilon} + \frac{2}{1+\epsilon} + 1 - [\epsilon] \right) \frac{1}{t} - (1+\epsilon) \frac{b}{\gamma_d^{1+\epsilon}} . \quad (31)$$

By employing (18) in the r.h.s. of (31) we obtain

$$g_1(t) \geq \bar{g}(t) \equiv t^\epsilon - \frac{R^2}{\gamma_d^2} \left(\frac{1+\epsilon}{2} t^\epsilon (\ln t)^{1-\epsilon} + \frac{3+\epsilon}{2} (1 - [\epsilon]) \right) \frac{1}{t} - (1+\epsilon) \frac{b}{\gamma_d^{1+\epsilon}} . \quad (32)$$

For $0 < \epsilon < 1$ $\bar{g}(t)$ becomes a function of the type considered in Lemma 4 with $\alpha = (2 + \epsilon) \frac{R^2}{\gamma_d^2} \geq 2 + \epsilon$. Obviously t_0 of Lemma 4 satisfies $g_1(t_0) \geq \bar{g}(t_0) > 0$ and according to (29) is an upper bound on t_{b_1} . Also for $\epsilon = 1$ $\bar{g}(t)$ becomes a function of the type considered in Lemma 1 and t_b of Lemma 1 is an upper bound on t_{b_1} . Actually in this very special case t_{b_1} coincides with t_b (since (32) holds as an equality) which, in turn, coincides with its upper and lower bound. This, given that $t_c \leq t_{b_1}$, completes the proof of (19).

Alternatively using $-\frac{1}{\epsilon} + \frac{2}{1+\epsilon} \leq 0$, $1 - [\epsilon] \leq (1 - [\epsilon])t^\epsilon$ and $(1+\epsilon)(1+\epsilon-[\epsilon]) = (1+\epsilon)^{2-[\epsilon]}$ in the r.h.s. of (31) we obtain

$$g_1(t) \geq \tilde{g}(t) \equiv t^\epsilon - \frac{1}{2\epsilon}(1+\epsilon)^{2-[\epsilon]} \frac{R^2}{\gamma_d^2} t^{\epsilon-1} - (1+\epsilon) \frac{b}{\gamma_d^{1+\epsilon}} .$$

The function $\tilde{g}(t)$ is of the type considered in Lemma 1 and its only root \tilde{t}_b satisfying

$$\tilde{g}(\tilde{t}_b) = 0 \quad (33)$$

is an upper bound on the number of updates looser than t_{b_1} i.e. $t_{b_1} \leq \tilde{t}_b$. Moreover, the upper bound on \tilde{t}_b from Lemma 1 is an alternative upper bound on t_c . Combining (30) for $N = 1$, the inequality $t_{b_1} \leq \tilde{t}_b$ and (33) we obtain

$$f \geq \frac{b}{\gamma_d^{1+\epsilon} t_{b_1}^\epsilon} \geq \frac{b}{\gamma_d^{1+\epsilon} \tilde{t}_b^\epsilon} = \left\{ \frac{R^{1+\epsilon}}{b} \frac{1}{2\epsilon} (1+\epsilon)^{2-[\epsilon]} \left(\frac{R}{\gamma_d} \right)^{1-\epsilon} \tilde{t}_b^{\epsilon-1} + 1 + \epsilon \right\}^{-1} . \quad (34)$$

Additionally, $\tilde{t}_{lb1} \equiv \frac{1}{2\epsilon}(1+\epsilon)^{2-[\epsilon]} \frac{R}{\gamma_d}$ is a lower bound on \tilde{t}_b since it is lower than the lower bound inferred from Lemma 1. Replacing \tilde{t}_b with its lower bound \tilde{t}_{lb1} in the r.h.s. of (34) we get the weaker bound (20). \square

Remark 3. The lower bounds (30) and (34) on the fraction f involving the unknown maximum margin γ_d are of great theoretical importance because they guarantee before running that the algorithm will achieve a margin which is a more substantial fraction of γ_d than the one inferred from (20). As a consequence, values of the parameter b smaller than the ones inferred from (20) suffice in order for the before-running lower bound on the fraction f to be close to its asymptotic value $(1+\epsilon)^{-1}$. This is quantified in the following theorem.

Theorem 3. *The ℓ -margitron with $0 < \epsilon \leq 1$ and b (at least as large as the one) given by*

$$\frac{b}{R^{1+\epsilon}} = \frac{(1+\epsilon)^{3\epsilon-1-[\epsilon]}}{(2\epsilon\delta)^\epsilon} \left(\frac{\gamma_d}{R} \right)^{1-\epsilon} \quad (35)$$

($\delta > 0$) converges in a finite number of updates to a solution hyperplane possessing directional margin γ'_d which is a fraction f of the maximum directional margin γ_d obeying the inequality

$$f = \frac{\gamma'_d}{\gamma_d} \geq (\delta + 1 + \epsilon)^{-1} . \quad (36)$$

Proof. Notice that

$$\tilde{t}_{lb2} \equiv \left((1+\epsilon) \frac{b}{\gamma_d^{1+\epsilon}} \right)^{\frac{1}{\epsilon}} = \frac{(1+\epsilon)^{3-[\epsilon]}}{2\epsilon\delta} \frac{R^2}{\gamma_d^2}$$

is a lower bound on \tilde{t}_b of (33) since it is lower than the lower bound inferred from Lemma 1. Replacing \tilde{t}_b with its lower bound \tilde{t}_{lb2} in the r.h.s. of (34) completes the proof. (Larger b 's may be regarded as corresponding to smaller δ 's.) \square

Remark 4. For $\epsilon \ll 1$ a more accurate determination of b ensuring that (36) holds is obtained from $\frac{b}{R^{1+\epsilon}} = \omega^\epsilon \left(\frac{\gamma_d}{R}\right)^{1-\epsilon}$ with $\omega = \frac{1}{\delta}(1-\epsilon)(1+e^{-1})(2+\epsilon)(1+\epsilon)^{\frac{\epsilon-1}{\epsilon}} \ln\left(\frac{1}{\delta}e^{\frac{1}{1-\epsilon}}(1-\epsilon)(1+e^{-1})(2+\epsilon)(1+\epsilon)^{\frac{\epsilon-1}{\epsilon}}\frac{R^2}{\gamma_d^2}\right)$ and $0 < \delta \leq e^{-1}(1+e^{-1})(2+\epsilon)\frac{R^2}{\gamma_d^2}$. For such a b and taking into account the constraint on δ it

can be verified that $\bar{t}_{\text{lb}} \equiv \left((1+\epsilon)\frac{b}{\gamma_d^{1+\epsilon}}\right)^{\frac{1}{\epsilon}} = (1+\epsilon)^{\frac{1}{\epsilon}}\omega\frac{R^2}{\gamma_d^2}$ satisfies the inequality $\bar{t}_{\text{lb}} > e$. Moreover, any possible root of $\bar{g}(t)$ defined in (32) and the single root t_{b_1} of $g_1(t)$ are necessarily larger than \bar{t}_{lb} . Therefore, since $\bar{t}_{\text{lb}} > e$ and given that $\frac{d\bar{g}}{dt} > 0$ ($\frac{d}{dt}\frac{\ln t}{t} < 0$) for $t > e$ there is a single root \bar{t}_b of $\bar{g}(t)$ satisfying $\bar{t}_b \geq t_{b_1} > \bar{t}_{\text{lb}} > e$. Combining (30) for $N = 1$ with the last inequality and the relation $\bar{g}(\bar{t}_b) = 0$ we get

$$\begin{aligned} f &\geq \frac{b}{\gamma_d^{1+\epsilon}t_{b_1}^\epsilon} \geq \frac{b}{\gamma_d^{1+\epsilon}\bar{t}_b^\epsilon} = \left\{\frac{R^2}{b\gamma_d^{1-\epsilon}}\left(\frac{1+\epsilon}{2}\bar{t}_b^\epsilon(\ln \bar{t}_b)^{1-\epsilon} + \frac{3+\epsilon}{2}\right)\bar{t}_b^{-1} + 1 + \epsilon\right\}^{-1} \\ &> \left\{(2+\epsilon)\frac{R^2}{b\gamma_d^{1-\epsilon}}\left(\frac{\ln \bar{t}_b}{\bar{t}_b}\right)^{1-\epsilon} + 1 + \epsilon\right\}^{-1} > \left\{(2+\epsilon)\frac{R^2}{b\gamma_d^{1-\epsilon}}\left(\frac{\ln \bar{t}_{\text{lb}}}{\bar{t}_{\text{lb}}}\right)^{1-\epsilon} + 1 + \epsilon\right\}^{-1}. \end{aligned}$$

Let $x = \frac{1}{\delta}e^{\frac{1}{1-\epsilon}}(1-\epsilon)(1+e^{-1})(2+\epsilon)(1+\epsilon)^{\frac{\epsilon-1}{\epsilon}}\frac{R^2}{\gamma_d^2}$. Then, $\omega\frac{R^2}{\gamma_d^2} = e^{-\frac{1}{1-\epsilon}}x \ln x$ and

$$\begin{aligned} (2+\epsilon)\frac{R^2}{b\gamma_d^{1-\epsilon}}\left(\frac{\ln \bar{t}_{\text{lb}}}{\bar{t}_{\text{lb}}}\right)^{1-\epsilon} &= (2+\epsilon)(1+\epsilon)^{\frac{\epsilon-1}{\epsilon}}\omega^{-1}\left(\frac{1}{\epsilon}\ln(1+\epsilon) + \ln\left(\omega\frac{R^2}{\gamma_d^2}\right)\right)^{1-\epsilon} \\ &< (2+\epsilon)(1+\epsilon)^{\frac{\epsilon-1}{\epsilon}}\omega^{-1}\left(1 + (1-\epsilon)\ln\left(\omega\frac{R^2}{\gamma_d^2}\right)\right) = \frac{\delta}{(1+e^{-1})}\frac{\ln(x \ln x)}{\ln x} \leq \delta \quad (37) \end{aligned}$$

($\ln \ln x / \ln x \leq e^{-1}$). Thus, our choice of b ensures that $f > (\delta + 1 + \epsilon)^{-1}$. Substituting b into (19) we conclude that in the ℓ -margintron as $\epsilon, \delta \rightarrow 0$ the upper bound on the number of updates $t_c \sim (\epsilon^{-1} + \delta^{-1} \ln \delta^{-1}) \ln(\epsilon^{-1} + \delta^{-1} \ln \delta^{-1}) R^2 / \gamma_d^2$. For $\epsilon \rightarrow 0$ with δ fixed, instead, the bound $\sim \epsilon^{-1} \ln \epsilon^{-1} R^2 / \gamma_d^2$. For $\delta \ll 1$ and $\delta/\epsilon < \lambda \approx 1$, however, a more accurate upper bound on t_c may be obtained by observing that

$$0 = \bar{g}(\bar{t}_b) > \bar{t}_b^\epsilon - (2+\epsilon)\frac{R^2}{\gamma_d^2}\left(\frac{\ln \bar{t}_b}{\bar{t}_b}\right)^{1-\epsilon} - \bar{t}_{\text{lb}}^\epsilon > \bar{t}_b^\epsilon - (2+\epsilon)\frac{R^2}{\gamma_d^2}\left(\frac{\ln \bar{t}_{\text{lb}}}{\bar{t}_{\text{lb}}}\right)^{1-\epsilon} - \bar{t}_{\text{lb}}^\epsilon$$

from where (using also (37))

$$\begin{aligned} \bar{t}_b &< \bar{t}_{\text{lb}}\left(1 + (2+\epsilon)\frac{R^2}{\gamma_d^2}\bar{t}_{\text{lb}}^{-\epsilon}\left(\frac{\ln \bar{t}_{\text{lb}}}{\bar{t}_{\text{lb}}}\right)^{1-\epsilon}\right)^{\frac{1}{\epsilon}} = \bar{t}_{\text{lb}}\left(1 + \frac{2+\epsilon}{1+\epsilon}\frac{R^2}{b\gamma_d^{1-\epsilon}}\left(\frac{\ln \bar{t}_{\text{lb}}}{\bar{t}_{\text{lb}}}\right)^{1-\epsilon}\right)^{\frac{1}{\epsilon}} \\ &< \bar{t}_{\text{lb}}\left(1 + \frac{\delta}{1+\epsilon}\right)^{\frac{1}{\epsilon}} < \bar{t}_{\text{lb}}(1+\delta)^{\frac{1}{\epsilon}} < \bar{t}_{\text{lb}}e^{\frac{\delta}{\epsilon}} = e^{\frac{\delta}{\epsilon}}(1+\epsilon)^{\frac{1}{\epsilon}}\omega\frac{R^2}{\gamma_d^2}. \end{aligned}$$

Taking into account that $t_c \leq t_{b_1} \leq \bar{t}_b$ we conclude that as $\delta \rightarrow 0$ with δ/ϵ bounded from above (e.g. $\delta = \epsilon \rightarrow 0$) the upper bound on $t_c \sim \delta^{-1} \ln \delta^{-1} R^2 / \gamma_d^2$.

Theorem 4. *There is a value of the parameter b for which the ℓ -margitron with $\epsilon \ll 1$ converges to a solution hyperplane with directional margin $\gamma'_d \geq (1 - 2\epsilon)\gamma_d$ in less than $\sim \epsilon^{-1} \ln \epsilon^{-1} R^2 / \gamma_d^2$ updates.*

Proof. Set $\delta = \epsilon$ in Remark 4 and notice that $f \geq (1 + 2\epsilon)^{-1} \geq 1 - 2\epsilon$. \square

Theorem 5. *Both the t - and the ℓ -margitron with $1 < \epsilon < 2$ converge in t_c updates, with t_c bounded from above by $\frac{R^2}{\gamma_d^2} + \left(\frac{2}{2-\epsilon} \frac{b}{\gamma_d^2}\right)^{\frac{1}{\epsilon}}$ and $\frac{R^2}{\gamma_d^2} + \left(\frac{2}{2-\epsilon} \frac{b}{\gamma_d^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}$ respectively, to a solution hyperplane possessing directional margin γ'_d which in the limit $b \rightarrow \infty$ satisfies the inequality $\gamma'_d \geq (1 - \frac{\epsilon}{2})\gamma_d$.*

Proof. For the t -margitron the analysis of Theorem 1 that led to (13) remains valid and the single root t_b of $g(t)$ still provides an upper bound on t_c . The bound on t_c stated in Theorem 5 is the upper bound on t_b inferred from Lemma 1. The analysis that led to (9) remains also valid but we are no longer allowed to replace t_c with its lower bound $t_c = 1$. Instead, we may replace t_c in (9) with its upper bound stated in Theorem 5. Then, as $b \rightarrow \infty$ we get $\gamma'_d \geq (1 - \frac{\epsilon}{2})\gamma_d$.

In the case of the ℓ -margitron $\mathbf{a}_t \cdot \mathbf{y}_k$ for a misclassified pattern \mathbf{y}_k may be bounded from above by employing (4) and (6) as $\mathbf{a}_t \cdot \mathbf{y}_k \leq b \|\mathbf{a}_t\|^{1-\epsilon} \leq b(\gamma_d t)^{1-\epsilon}$. Then, the analysis of Theorem 1 that led to (13) remains valid with the replacement of b by $b\gamma_d^{1-\epsilon}$. The bound on t_c stated in Theorem 5 is the upper bound on t_b inferred from Lemma 1. For the fraction f , instead, employing (4), (5), (11) and (13), with the last two relations taken at $t = t_b$ as equalities, we have

$$f = \frac{\gamma'_d}{\gamma_d} \geq \frac{b}{\gamma_d \|\mathbf{a}_{t_c}\|^\epsilon} \geq \frac{b}{\gamma_d^{1+\epsilon} t_b^\epsilon} = \left(\frac{R^2}{b \gamma_d^{1-\epsilon} t_b^{\epsilon-1}} + \frac{2}{2-\epsilon} \right)^{-1}.$$

Replacing t_b with its upper bound $\frac{R^2}{\gamma_d^2} + \left(\frac{2}{2-\epsilon} \frac{b}{\gamma_d^{1+\epsilon}}\right)^{\frac{1}{\epsilon}}$ in the above relation leads to a weaker bound from where we get $\lim_{b \rightarrow \infty} f \geq 1 - \frac{\epsilon}{2}$. \square

4 Experiments

To reduce the computational cost we follow [17] and form a reduced “active set” of patterns consisting of the ones found misclassified during each epoch which are then cyclically presented to the Margitron algorithm for N_{ep} mini-epochs unless no update occurs during a mini-epoch. Subsequently, a new full epoch involving all the patterns takes place giving rise to a new active set. The algorithm terminates only if no mistake occurs during a full epoch. This procedure clearly amounts to a different way of sequentially presenting the patterns to the algorithm and does not affect the applicability of our theoretical analysis.

We compare the t - and the ℓ -margitron with SVMs on the basis of their ability to achieve fast convergence to a certain approximation of the “optimal” hyperplane in the feature space where the patterns are linearly separable. For

Table 1. Results of a comparative study of SVM^l, t -margitron and ℓ -margitron.

data set	Δ	SVM ^l $\epsilon = 0.01$		ρ	N_{ep}	t - margitron				ℓ - margitron			
		$10^3\gamma'$	Secs			ϵ	$\frac{10^8 b}{R^2}$	$10^3\gamma'$	Secs	ϵ	$\frac{10^8 b}{R^{1+\epsilon}}$	$10^3\gamma'$	Secs
Adult	1	8.4899	1810.3	0	50	0.001	491.1	8.4917	72.2	0.0005	220.4	8.4903	68.2
Web	1	20.941	250.0	0.1	10	0.2	8400	20.944	17.6	0.2	1250	20.942	17.4
C11	0.1	1.7818	6172.0	0.1	50	0.2	10000	1.7822	631.7	0.2	1600	1.7821	655.3
CCAT	0.1	0.9016	48235.0	0.1	50	0.1	514.2	0.9016	2324.8	0.1	285	0.9018	2369.7
Cover	10	15.774	47987.7	1	20	0.01	158.6	15.774	1866.1	0.005	121.7	15.776	1760.0

linearly separable data the feature space is the initial instance space whereas for linearly inseparable data (which is the case here) a space extended by as many dimensions as the instances is considered where each instance is placed at a distance Δ from the origin in the corresponding dimension. The extension generates a margin of at least Δ/\sqrt{n} with n being the number of patterns and amounts to adding a term Δ^2 to the diagonal entries of the kernel (linear in our case). Moreover, its employment is justified by the well-known equivalence between the hard margin optimisation in the extended space and the soft margin optimisation in the initial instance space with objective function $\|\mathbf{w}\|^2 + \Delta^{-2} \sum_i \xi_i^2$ involving the weight vector \mathbf{w} and the 2-norm of the slacks ξ_i [1]. We emphasize that SVMs and the Margitron are required to solve identical hard margin problems.

In our experiments SVMs are represented by SVM^{light} [5], denoted here as SVM^l, a decomposition method algorithm which is many orders of magnitude faster than standard SVMs. For SVM^l we choose a memory parameter $m = 400\text{MB}$ and a 1-norm soft margin parameter $C = 10^5$ (approximating $C = \infty$) since we are dealing with a hard margin problem in the appropriate feature space. The choice of the accuracy ϵ depends on the case. For the remaining parameters default values are used. The experiments were conducted on a 1.8 GHz Intel Pentium M processor with 504 MB RAM running Windows XP. The codes written in C++ were run using Microsoft's Visual C++ 5.0 compiler.

The datasets we used for training are the Adult (32561 instances, 123 binary attributes) and Web (49749 instances, 300 binary attributes) UCI datasets as compiled by Platt (see [5]), the test0 set from the Reuters RCV1 collection (199328 instances, 47236 attributes with average sparsity 0.16%) obtainable from http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm and the multiclass Covertypes (Cover) UCI dataset (581012 instances, 54 attributes). In the case of the RCV1 we considered both the C11 and the CCAT binary text classification tasks while in the case of the Covertypes dataset we studied the binary classification problem of the first class versus all the others. The Covertypes dataset was rescaled by multiplying all the attributes with 0.001.

In Table 1 we present the results (i.e. geometric margin γ' achieved and CPU secs needed) of our first comparative study involving the algorithms SVM^l, t -margitron and ℓ -margitron together with the values of the parameters employed. A solution hyperplane in the extended space was first obtained using SVM^l and subsequently the Margitron was required to obtain a solution of comparable

geometric margin. The extended space parameter Δ refers to both SVM^l and the Margitron while the augmented space parameter ρ and the number of mini-epochs N_{ep} only to the Margitron. Also, for the Margitron γ' is the geometric margin in the original (non-augmented) feature space with the augmentation providing for the bias. We see that the Margitron is at least 10-20 times faster than SVM^l on these rather large datasets. It is understood, of course, that some additional computer time was spent to locate the appropriate value of b .

Recently SVM^{perf} [6], a cutting-plane algorithm for training linear SVMs, was presented. We did make an attempt at including SVM^{perf} in our comparative study but we found that it requires a much longer CPU time to converge compared to SVM^l without even achieving as large values of the margin γ' . Table 2 contains our experimental results on the datasets Adult and Web ($\Delta = 1$). Apparently, the “accuracy” ϵ of SVM^{perf} is not directly related to the fraction of the maximum margin achieved.

Table 2. Results of experiments with SVM^{perf}.

data set	SVM ^{perf}			
	ϵ	C	$10^3\gamma'$	Secs
Adult	3×10^{-4}	10^8	5.9436	54450.3
Web	2×10^{-5}	10^8	20.891	7297.9

In Table 3 we present the directional margin γ'_d achieved by the t - and the ℓ -margitron together with the after-running estimate f_{est} of the ratio γ'_d/γ_d and its asymptotic value for comparison. Let us accept that the geometric margin γ' reported in Table 1 is larger than 99% of the maximum geometric margin γ as the accuracy $\epsilon = 0.01$ of SVM^l suggests. Then, taking into account that $\gamma \geq \gamma_d$ and that $(\gamma' - \gamma'_d)/\gamma' < 0.02$ we see that $\gamma'_d/\gamma_d > 0.97$. Thus, we may conclude that the estimates of Table 3 are certainly impressive given that they come from worst-case bounds which are not expected to be very tight and that they cannot, of course, exceed their asymptotic values.

Table 3. The directional margin γ'_d achieved by the t - and the ℓ -margitron together with the after-running estimate f_{est} of the ratio γ'_d/γ_d and its asymptotic value.

data set	t – margitron			ℓ – margitron		
	$10^3\gamma'_d$	f_{est}	$1 - \frac{\epsilon}{2}$	$10^3\gamma'_d$	f_{est}	$(1 + \epsilon)^{-1}$
Adult	8.4917	0.9898	0.9995	8.4903	0.9420	0.9995
Web	20.574	0.8645	0.9000	20.573	0.7561	0.8333
C11	1.7789	0.8923	0.9000	1.7787	0.8156	0.8333
CCAT	0.9016	0.9404	0.9500	0.9018	0.8752	0.9091
Cover	15.714	0.9873	0.9950	15.716	0.9703	0.9950

Table 4. A comparison between the ℓ -margitron (successive runnings) and SVM^l.

data set	$\ell - \text{margitron}$								SVM ^l		
	$\epsilon = 1$			$\epsilon = 0.1$							
	$10^3 \gamma'_d$	$10^3 \gamma_d^{\text{up}}$	Secs	$\frac{10^3 b}{R^{1+\epsilon}}$	$10^3 \gamma'_d$	f_{est}	$10^3 \gamma'$	Secs	ϵ	$10^3 \gamma'$	Secs
Adult	6.8839	11.352	3.2	577	8.3274	0.838	8.3274	39.5	0.055	8.3257	1178.9
Web	19.202	29.840	5.0	551	20.677	0.864	21.053	44.0	0.0031	21.051	291.2
C11	1.5435	2.5607	121.7	506	1.7765	0.863	1.7798	774.7	0.012	1.7798	5952.0
CCAT	0.7800	1.2701	366.7	270	0.8989	0.855	0.8989	1771.5	0.0165	0.8984	42548.8
Cover	10.644	19.566	334.5	301	14.674	0.816	14.735	527.4	0.085	14.718	29402.8

From (16) and (35) it becomes apparent that the minimal value of b guaranteeing the desired accuracy depends on the maximum directional margin γ_d . Moreover, this dependence becomes increasingly crucial with decreasing ϵ . This last observation prompts us to proceed to a determination of the large margin solution in successive runnings starting with the more insensitive to the value of γ_d Margitron with $\epsilon = 1$ and gradually moving towards employing algorithms with smaller ϵ 's able to guarantee larger fractions of γ_d . Each running in this process will provide us with an interval in which the value of γ_d lies which, hopefully, will shrink as we move towards smaller ϵ 's. This information will then allow us to fix the value of b to be used in the next running. The lower bound on γ_d will be the margin γ'_d achieved. The upper bound γ_d^{up} will be provided by exploiting the after-running estimate f_{est} of γ'_d/γ_d which gives $\gamma_d^{\text{up}} = \gamma'_d f_{\text{est}}^{-1}$. Alternatively, we may employ the upper bound on the number of updates t_c required for convergence to obtain a value for γ_d^{up} . For $\epsilon = 1$ this gives $\gamma_d^{\text{up}} = R\sqrt{(1 + 2b/R^2)} t_c^{-1}$ which is usually lower than the upper bound $(R^2/b + 2) \gamma'_d$ on γ_d obtained from $\gamma'_d/\gamma_d \geq (R^2/b + 2)^{-1}$. This procedure may be followed using either the t - or the ℓ -margitron but in the former case we may encounter difficulties for $\epsilon < 1/2$ due to the lack of the strong before-running guarantees stemming from (15).

In Table 4 we present the results of a second comparative study between the ℓ -margitron and SVM^l. For the ℓ -margitron we followed the procedure of successive runnings that we just described involving only two stages with ϵ values 1 and 0.1. The extended and augmented feature spaces were identical to the ones of Table 1 and a common value $N_{\text{ep}} = 50$ was chosen for all datasets. Also, in the first stage ($\epsilon = 1$) we made the common choice $b/R^2 = 5$ and obtained γ_d^{up} from the relation $\gamma_d^{\text{up}} = R\sqrt{11t_c^{-1}}$. Then, in the second stage ($\epsilon = 0.1$) we fixed b from (35) with $\delta = (\gamma_d/\gamma_d^{\text{up}})^{\frac{1-\epsilon}{\epsilon}} \leq 1$ (which eliminates the dependence of b on γ_d) employing the γ_d^{up} obtained in the first stage. This way we shift the uncertainty in $\gamma_d/\gamma_d^{\text{up}}$ to the before-running accuracy δ and rely on the after-running lower bound f_{est} on γ'_d/γ_d to assess the accuracy actually achieved. We see that f_{est} is well above 0.8 for all datasets. A comparison with SVM^l on solutions of comparable margin reveals that the ℓ -margitron remains considerably faster even if the time spent to fix b is taken into account.

5 Conclusions

We generalised the classical Perceptron algorithm with margin by constructing the Margitron, a family of incremental large margin classifiers all the members of which employ the original perceptron update. The Margitron consists of two classes, namely the t -margitron with algorithms involving explicitly the number of updates and the ℓ -margitron the members of which depend only on the length of the weight vector and as such lie closer in spirit to the Perceptron. We proved that as the parameter ϵ decreases from 2 to 0 the corresponding algorithms in both classes converge in a finite number of updates to hyperplanes possessing a guaranteed fraction of the maximum margin the largest possible value of which varies continuously in the interval $(0, 1)$. The Perceptron with margin belongs to both classes and is associated with the middle point of the above intervals. Finally, our experimental comparative study between algorithms from the margitron family and SVM^{light} on tasks involving linear kernels and 2-norm soft margin revealed that the Margitron is a serious alternative to linear SVMs.

References

1. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines (2000) Cambridge, UK: Cambridge University Press
2. Duda, R.O., Hart, P.E.: Pattern classification and scene analysis (1973) Wiley
3. Gentile, C.: A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research* **2** (2001) 213–242
4. Gradshteyn, I.S., Ryzhik, I.M.: Tables of integrals, series and products (2007) Academic Press
5. Joachims, T.: Making large-scale svm learning practical. In *Advances in kernel methods-support vector learning* (1999) MIT Press
6. Joachims, T.: Training linear svms in linear time. *KDD* (2006) 217–226
7. Kivinen, J., Smola, A.J., Williamson, R.C.: Online learning with kernels. *IEEE TSP*, **52** (2004) 2165–2176
8. Krauth, W., Mézard, M.: Learning algorithms with optimal stability in neural networks. *Journal of Physics* **A20** (1987) L745–L752
9. Li, Y., Long, P.: The relaxed online maximum margin algorithm. *Machine Learning*, **46** (2002) 361–387
10. Li, Y., Zaragoza, H., Herbrich, R., Shawe-Taylor, J., Kandola, J.: The perceptron algorithm with uneven margins. *ICML* (2002) 379–386
11. Novikoff, A.B.J.: On convergence proofs on perceptrons. In *Proc. Symp. Math. Theory Automata*, Vol. 12 (1962) 615–622
12. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65** (6) (1958) 386–408
13. Shalev-Schwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for svm. *ICML* (2007) 807–814
14. Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., Anthony, M.: Structural risk minimization over data-dependent hierarchies. *IEEE TIT*, **44**(5) (1998) 1926–1940
15. Tsampouka, P., Shawe-Taylor, J.: Analysis of generic perceptron-like large margin classifiers. *ECML* (2005) 750–758

16. Tsampouka, P., Shawe-Taylor, J.: Constant rate approximate maximum margin algorithms. ECML (2006) 437–448
17. Tsampouka, P., Shawe-Taylor, J.: Approximate maximum margin algorithms with rules controlled by the number of mistakes. ICML (2007) 903–910
18. Vapnik, V.: Statistical learning theory (1998) Wiley