A Feedback Analysis of Perceptron Learning for Neural Networks*

A. H. Sayed
Dept. Electrical and Computer Eng.
University of California
Santa Barbara, CA 93106

M. Rupp
Wireless Technology, Research Dept.
AT&T
791 Holmdel-Keyport Rd.
Holmdel NJ 07733–0400

Abstract

This paper provides a time-domain feedback analysis of the Perceptron Learning Algorithm. It studies the robustness performance of the algorithm in the presence of uncertainties that might be due to noisy perturbations in the reference signal or to modeling mismatch. In particular, bounds are established on the step-size parameter in order to guarantee that the resulting algorithm will behave as a robust filter in the sense of H^∞–theory. The paper also establishes that an intrinsic feedback structure can be associated with the training scheme. The feedback configuration is motivated via energy arguments and is shown to consist of two major blocks: a time-variant lossless (i.e., energy preserving) feedforward path and a time-variant feedback path. The stability of the feedback structure is then analyzed via the small gain theorem and choices for the step-size parameter in order to guarantee faster convergence are further derived by appealing to the mean-value theorem. Simulation results are included to validate the findings.

1 Introduction

Applications of neural networks span a variety of areas in pattern recognition, filtering, and control. When supervised learning is employed, a training phase is always necessary. During this phase, a recursive update procedure is used to estimate the weight vector of the linear combiner that "best" fits the given data [1, 3, 5]. The recursive procedure often requires that a suitable adaptation gain (or step-size parameter) be chosen and, in most cases, heuristics and trial-and-error experiences are used to select a suitable step-size value for the training period.

The "common" practice is to choose small adaptation gains. But the smaller the adaptation gain the slower the convergence speed. In several cases, especially in large-scale applications with many weights and many training patterns, this may require a considerable amount of time and machine power.

In recent work on the robustness analysis of adaptive schemes [6], we have addressed the following two issues:

1. We have shown how to select the adaptation gain in order to guarantee a robust behaviour in the presence of noise and modeling uncertainties.
2. We have also shown how to select the adaptation gain in order to guarantee faster convergence.

Our formulation highlights an intrinsic feedback structure for most adaptive schemes and combines tools from system theory, control, and signal processing such as: state-space descriptions, feedback analysis and the small gain theorem, H^∞–tools, and transmission lines and lossless systems.

In this paper we focus on the so-called Perceptron Learning Algorithm (PLA, for short), which involves a nonlinear functional in the update equation due to the presence of an activation function (usually a sigmoid function). We show how to extend the feedback arguments of [6] in order to handle the presence of the nonlinearity. In particular, we also establish the existence of a feedback structure that can be associated with the training algorithm.

The feedback structure provides physical insights into the energy propagation as the algorithm progresses in time. This enables us to suggest modifications to the training algorithm, in terms of selections of the adaptation gain, in order to accelerate the convergence speed during the training phase.

Notation. We use small boldface letters to denote vectors (e.g., $\mathbf{u}$), "$*$" to denote Hermitian conjugation, and $\|\mathbf{x}\|_2$ to denote the Euclidean norm of a vector. We also use subscripts for time-indexing of vector quantities (e.g., $\mathbf{u}_i$) and parenthesis for time-indexing of scalar quantities (e.g., $v(i)$). All vectors are column vectors except for the row vectors $\mathbf{u}_i$.

2 The Perceptron

Consider two sets, $\mathcal{S}_0$ and $\mathcal{S}_1$, of M–dimensional complex-valued row vectors $\mathbf{u}$ that are characterized by either property A or property B. If the two sets are linearly separable, then a classification scheme that can be used to decide whether a given vector $\mathbf{u}$ belongs to one class or the other is to employ a perceptron device [1, 3, 5].

The perceptron consists of a linear combiner, whose column weight vector we denote by $\mathbf{w}$, followed by a nonlinearity $f(z)$ (also known as an activation function), as depicted in Figure 1. A common choice for

*This work was supported in part by the National Science Foundation under Award No. MIP-9409319.

1058-6393/96 $5.00 © 1996 IEEE
Proceedings of ASILOMAR-29

$f(z)$ is the sigmoid function

$$f_\beta(z) = \frac{1}{1+e^{-\beta z}}, \quad \beta > 0. \tag{1}$$

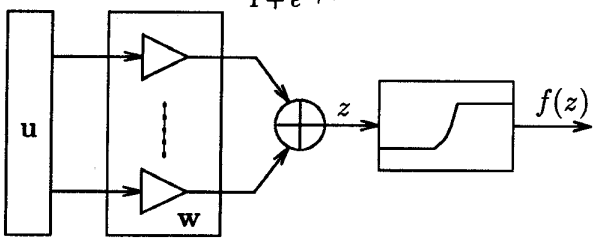

Figure 1: *The Perceptron structure.*

But, more generally, it can be any monotonically increasing function. The outcome $f(z)$ can be interpreted as the likelihood that the input vector belongs to $\mathcal{S}_0$ or $\mathcal{S}_1$.

3 The Perceptron Learning Algorithm

Consider a collection of input vectors $\{\mathbf{u}_i\}$ with the corresponding (desired) output (or reference) values $\{y(i)\}$. The $\{y(i)\}$ are assumed to belong to the range of the activation function $f(\cdot)$, i.e.,

$$y(i) = f(\mathbf{u}_i\mathbf{w}) \quad \text{for some } \mathbf{w}. \tag{2}$$

This is in agreement with the models and assumptions used in [2, 7].

In supervised learning, the Perceptron is presented with the given input-output data $\{\mathbf{u}_i, y(i)\}$ and the objective is to estimate $\mathbf{w}$. The Perceptron Learning Algorithm computes recursive estimates of $\mathbf{w}$ as follows (using an initial guess $\mathbf{w}_{-1}$):

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu\mathbf{u}_i^* [y(i) - f(\mathbf{u}_i\mathbf{w}_{i-1})]. \tag{3}$$

For generality, we consider in this paper the possibility of noisy perturbations in the reference signal $y(i)$. These can be due to model mismatching or to measurement noise. We denote the perturbed references by $\{d(i)\}$ (which are now the given data instead of $\{y(i)\}$), say

$$d(i) = f(\mathbf{u}_i\mathbf{w}) + v(i) = y(i) + v(i), \tag{4}$$

where $v(i)$ denotes the noise term. Correspondingly, we study the following general form of recursion (3):

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu(i)\mathbf{u}_i^* [d(i) - f(\mathbf{u}_i\mathbf{w}_{i-1})], \tag{5}$$

where $d(i)$ replaces $y(i)$ and where we have allowed for a time-variant step-size parameter $\mu(i)$.

3.1 Error Measures

The following error quantities are useful for our later analysis: $\tilde{\mathbf{w}}_i$ denotes the difference between the true weight $\mathbf{w}$ and its estimate $\mathbf{w}_i$, $\tilde{\mathbf{w}}_i = \mathbf{w} - \mathbf{w}_i$, $e_a(i)$ denotes the *a priori* estimation error, $e_a(i) = \mathbf{u}_i\tilde{\mathbf{w}}_{i-1} = z(i) - \hat{z}(i)$, and $e_p(i)$ denotes the *a posteriori* estimation error, $e_p(i) = \mathbf{u}_i\tilde{\mathbf{w}}_i$. It follows from (5) that the weight-error vector satisfies the recursion

$$\tilde{\mathbf{w}}_i = \tilde{\mathbf{w}}_{i-1} - \mu(i)\mathbf{u}_i^* [d(i) - f(\mathbf{u}_i\mathbf{w}_{i-1})]. \tag{6}$$

3.2 Robustness Issues

In the sequel we focus on model (4) and study the robustness behaviour of the update recursion (5).

Intuitively, a robust algorithm is one for which the estimation errors are consistent with the disturbances in the sense that "small" disturbances would lead to "small" estimation errors, no matter what the disturbances are! This is not generally true for any adaptive filter: the estimation errors can still be large even in the presence of small disturbances.

The robustness issue is addressed here in a purely deterministic framework and without assuming prior knowledge of noise statistics. This is especially useful in situations where prior statistical information is missing since a robust design would guarantee a desired level of robustness independent of the noise statistics. In loose terms, robustness would imply that the ratio of the estimation error energy to the noise or disturbance energy will be guaranteed to be upper bounded by a positive constant, say the constant one,

$$\frac{\text{estimation error energy}}{\text{disturbance energy}} \leq 1. \tag{7}$$

From a practical point of view, a relation of the form (7) is desirable since it guarantees that the resulting estimation error energy will be upper bounded by the disturbance energy, no matter what the nature and the statistics of the disturbances are. One of the contributions of this work is to show how to select the adaptation gains $\mu(i)$ in (5) in order to guarantee i) a robust behavior and ii) faster convergence. This is addressed in the next sections.

4 A Contractive Mapping

We denote the difference $[d(i) - f(\mathbf{u}_i\mathbf{w}_{i-1})]$ in (6) by $\tilde{e}_a(i)$ and note that it is equal to $[e_a(i)+\tilde{v}(i)]$, where the modified disturbance $\tilde{v}(i)$ is defined by:

$$\tilde{v}(i) = -e_a(i) + f(\mathbf{u}_i\mathbf{w}) - f(\mathbf{u}_i\mathbf{w}_{i-1}) + v(i). \tag{8}$$

This allows us to rewrite (6) as

$$\tilde{\mathbf{w}}_i = \tilde{\mathbf{w}}_{i-1} - \mu(i)\mathbf{u}_i^*\tilde{e}_a(i). \tag{9}$$

If we now compute the squared norm (i.e., energies) of both sides of (9), we conclude that the following equality always holds,

$$\|\tilde{\mathbf{w}}_i\|_2^2 + \mu(i)\,|e_a(i)|^2 + \mu(i)\,(1 - \mu(i)\|\mathbf{u}_i\|_2^2)\,|\tilde{e}_a(i)|^2 = \|\tilde{\mathbf{w}}_{i-1}\|_2^2 + \mu(i)\,|\tilde{v}(i)|^2.$$

This equality allows us to conclude that the following energy bounds are always satisfied, where we have introduced the parameter $\bar{\mu}(i) = 1/\|\mathbf{u}_i\|_2^2$.

Lemma 1 *Consider the perceptron learning recursion (3)-(4). It always holds, at each time instant i, that*

$$\frac{\|\tilde{\mathbf{w}}_i\|_2^2 + \mu(i)\,|e_a(i)|^2}{\|\tilde{\mathbf{w}}_{i-1}\|_2^2 + \mu(i)\,|\tilde{v}(i)|^2} \begin{cases} \leq 1 & \text{for } 0<\mu(i)<\bar{\mu}(i) \\ = 1 & \text{if } \mu(i) = \bar{\mu}(i) \\ \geq 1 & \text{for } \mu(i) > \bar{\mu}(i) \end{cases}$$

where $\tilde{v}(i)$ is the modified disturbance given by (8).

The first two inequalities in the statement of the lemma establish that if the adaptation gain is chosen such that $\mu(i) \leq \bar{\mu}(i)$, then the mapping from the signals $\{\tilde{\mathbf{w}}_{i-1}, \sqrt{\mu(i)}\tilde{v}(i)\}$ to the signals $\{\tilde{\mathbf{w}}_i, \sqrt{\mu(i)}e_a(i)\}$ is contractive. [A linear map that takes x to y, say $y = T[x]$, is said to be contractive if for all x we have $\|T[x]\|_2^2 \leq \|x\|_2^2$. That is, the output energy does not exceed the input energy].

But since this contractivity property holds for each time instant i, it should also hold globally over an interval of time. Indeed, assuming $\mu(i) \leq \bar{\mu}(i)$ over $0 \leq i \leq N$, it follows from Lemma 1 that

$$\|\tilde{\mathbf{w}}_N\|_2^2 + \sum_{i=0}^{N} \mu(i)\,|e_a(i)|^2 \leq \|\tilde{\mathbf{w}}_{-1}\|_2^2 + \sum_{i=0}^{N} \mu(i)|\tilde{v}(i)|^2.$$

5 A Feedback Structure

The bounds of Lemma 1 can be described in an alternative form that leads to an interesting feedback structure. For this purpose, we first note that it can also be shown that the update equation (5) can be written in the form (cf. the analysis in [6]):

$$\mathbf{w}_i \;=\; \mathbf{w}_{i-1} + \bar{\mu}(i)\mathbf{u}_i^*[e_a(i) - e_p(i)]\,, \qquad (10)$$

where we have used the fact that

$$e_p(i) = e_a(i) - \frac{\mu(i)}{\bar{\mu}(i)}\,[f(\mathbf{u}_i\mathbf{w}) - f(\mathbf{u}_i\mathbf{w}_{i-1}) + v(i)]\,. \qquad (11)$$

Consequently,

$$\tilde{\mathbf{w}}_i = \tilde{\mathbf{w}}_{i-1} - \bar{\mu}(i)\mathbf{u}_i^*[e_a(i) - e_p(i)] \qquad (12)$$

Relation (12) has the same form as the update (9), except for a different disturbance ($\tilde{v}(i)$ is now replaced by $-e_p(i)$) and for a step-size that is equal to $\bar{\mu}(i)$ itself. Hence, the same arguments that led to Lemma 1 would imply that the following equality holds for all possible choices of $\mu(i)$:

$$\frac{\|\tilde{\mathbf{w}}_i\|_2^2 + \bar{\mu}(i)\,|e_a(i)|^2}{\|\tilde{\mathbf{w}}_{i-1}\|_2^2 + \bar{\mu}(i)\,|e_p(i)|^2} = 1. \qquad (13)$$

This establishes a lossless mapping $\overline{T}_i$ from $\{\tilde{\mathbf{w}}_{i-1}, \sqrt{\bar{\mu}(i)}e_p(i)\}$ to $\{\tilde{\mathbf{w}}_i, \sqrt{\bar{\mu}(i)}e_a(i)\}$.

If we further apply the mean-value theorem to the activation function $f(z)$, we can write

$$f(\mathbf{u}_i\mathbf{w}) - f(\mathbf{u}_i\mathbf{w}_{i-1}) = f'(\eta(i))e_a(i),$$

for some point $\eta(i)$ along the segment connecting $\mathbf{u}_i\mathbf{w}$ and $\mathbf{u}_i\mathbf{w}_{i-1}$. Therefore, (11) leads to

$$-\bar{\mu}^{\frac{1}{2}}(i)e_p(i) = \frac{\mu(i)}{\bar{\mu}^{\frac{1}{2}}(i)}v(i) - \left[1 - f'(\eta)\frac{\mu(i)}{\bar{\mu}(i)}\right]\bar{\mu}^{\frac{1}{2}}(i)e_a(i).$$

This shows that the overall mapping from the *original* (weighted) disturbances $\sqrt{\bar{\mu}(\cdot)}v(\cdot)$ to the resulting

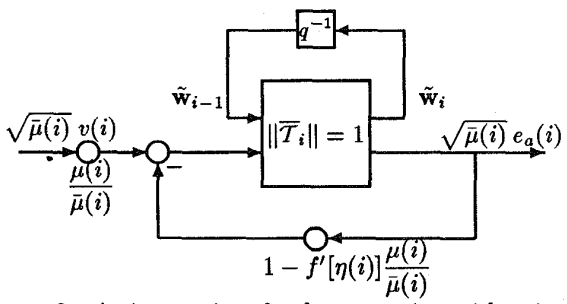

Figure 2: *A time-variant lossless mapping with gain feedback for the perceptron learning algorithm.*

a priori (weighted) estimation errors $\sqrt{\bar{\mu}(\cdot)}e_a(\cdot)$ can be expressed in terms of a feedback structure, as shown in Figure 2.

The stability of such feedback structures can be studied via tools that are by now standard in system theory (e.g., the small gain theorem [4]).

6 l_2–Stability

Define $\gamma(N) = \max_{0\leq i\leq N}\ \mu(i)/\bar{\mu}(i)$ and

$$\Delta(N) \;\triangleq\; \max_{0\leq i\leq N} \left|1 - f'(\eta(i))\frac{\mu(i)}{\bar{\mu}(i)}\right|$$

That is, $\Delta(N)$ is the maximum absolute value of the gain of the feedback loop over $0 \leq i \leq N$.

It can be shown that if $\Delta(N) < 1$ (e.g., [6]) then

$$\sqrt{\sum_{i=0}^{N} \bar{\mu}(i)\,|e_a(i)|^2} \;\leq\; \frac{1}{1-\Delta(N)} \left[\,\|\tilde{\mathbf{w}}_{-1}\|_2 + \gamma(N)\sqrt{\sum_{i=0}^{N} \bar{\mu}(i)|v(i)|^2}\,\right]. \qquad (14)$$

Expression (14) compares the energies of a priori estimation errors and the disturbances (but now normalized by $\bar{\mu}(i)$ rather than $\mu(i)$). In particular, it establishes that the map from $\{\tilde{\mathbf{w}}_{-1}, \sqrt{\bar{\mu}(\cdot)}v(\cdot)\}$ to $\{\sqrt{\bar{\mu}(\cdot)}e_a(\cdot)\}$ is l_2–stable (it maps a finite energy sequence to another finite energy sequence) [4]. The condition $\Delta(N) < 1$ can be interpreted as a manifestation of the small gain theorem [4]. It is also straightforward to verify that $\Delta(N) < 1$ means that $\mu(i)$ should satisfy

$$0 < \mu(i)f'(\eta(i)) < 2\bar{\mu}(i) = \frac{2}{\|\mathbf{u}_i\|_2^2}\,. \qquad (15)$$

7 Energy Propagation

The flow of energy through the feedback connection of Figure 2 provides further insights into the convergence behaviour of the training algorithm. For this purpose, let us ignore the measurement noise $v(i)$.

If $\mu(i)$ is such that $\mu(i)f'(\eta(i)) = \bar{\mu}(i)$, then the feedback loop is disconnected. This means that there

is no energy flowing back into the lower input of the lossless section from its lower output $e_a(\cdot)$. The losslessness of the feedforward path then implies that $E_w(i) = E_w(i-1) - E_e(i)$, where we are denoting by $E_e(i)$ the energy of $\sqrt{\bar{\mu}(i)}\, e_a(i)$ and by $E_w(i)$ the energy of $\tilde{\mathbf{w}}_i$.

But what if $\mu(i)f'(\eta(i)) \neq \bar{\mu}(i)$? In this case the feedback path is active and the convergence speed will be affected (becomes slower). Indeed, we now have

$$E_w(i) = E_w(i-1) - \underbrace{\left(1 - \left|1 - f'(\eta(i))\frac{\mu(i)}{\bar{\mu}(i)}\right|^2\right)}_{\tau(i)} E_e(i),$$

where we have defined the coefficient $\tau(i)$. It is easy to verify that as long as $\mu(i)f'(\eta(i)) \neq \bar{\mu}(i)$ we always have $0 < \tau(i) < 1$. That is, $\tau(i)$ is strictly less than one and the rate of decrease in the energy of $\tilde{\mathbf{w}}_i$ is lowered.

8 Optimal Choices of Step-Sizes

The energy arguments suggest that faster convergence occurs when $\mu(i)$ is chosen such that $\mu(i) = \bar{\mu}(i)/f'(\eta(i))$ (which is the middle point of the interval suggested by (15)). But $\eta(i)$ is still unknown and we therefore need to come up with suitable approximations.

The first (but not the most suitable) choice that comes to mind is to assume an upper bound on $f'(\cdot)$, say $f'(\eta) \leq f'_{\max}$ for all η. Then condition (15) can be replaced by the conservative requirement

$$0 < \mu(i) < 2/(f'_{\max}\|\mathbf{u}_i\|_2^2). \tag{16}$$

For a large bound f'_{max}, this condition can lead to small step-sizes and, hence, to slow convergence. For the commonly used activation functions, the maximum value of the derivative occurs at the origin. For example, for the sigmoid function we obtain $f'(0) = \beta/4$. We can therefore take $f_{\max} = \beta/4$ and choose the step-size parameter $\mu(i)$ according to $0 < \mu(i) < 8/(\beta\|\mathbf{u}_i\|_2^2)$. This is the same bound suggested in [2]. For improved convergence one might then be tempted to employ $\mu(i) = 4/(\beta\|\mathbf{u}_i\|_2^2)$. However, this value is very conservative and usually leads to unsatisfactory results, as the simulations at the end of this paper demonstrate.

For this reason, we take here an alternative route that avoids upper-bounding the derivative of the activation function. Instead, we provide good estimates for the instantaneous derivatives $f'(\eta)$.

To begin with, recall that $f'(\eta)$ is defined by $f'(\eta) = [f(z) - f(\mathbf{u}_i\mathbf{w}_{i-1})]/[z - \mathbf{u}_i\mathbf{w}_{i-1}]$, where $z = \mathbf{u}_i\mathbf{w}$. Unfortunately, z and $f(z)$ are not available since $\mathbf{w}$ itself is not known. But one possibility to proceed here is to employ $d(i)$ as an estimate for $f(z)$ since $d(i) = f(z) + v(i)$. This is especially useful if the reference sequence is noise-free or if the noise itself is sufficiently small. Now, with a "known" $f(z)$, it becomes possible to solve for z. This motivates us to suggest the following expression for the optimal step-size parameter (we refer to this construction as method A):

$$\mu_{opt}(i) = \bar{\mu}(i)\min\left(-\frac{\frac{\ln[f^{-1}(d(i))-1]}{\beta} + \mathbf{u}_i\mathbf{w}_{i-1}}{d(i) - f(\mathbf{u}_i\mathbf{w}_{i-1})}, T\right), \tag{17}$$

where T is used as a threshold value in order to prevent large step-sizes. This construction of the step-size requires the evaluation of a logarithm at each step.

An alternative procedure is to approximate $f'(\eta(i))$ by the average of $f'(z)$ (or $\approx f'(d(i))$) and $f'(\mathbf{u}_i\mathbf{w}_{i-1})$. This is a convenient approximation in light of the "piecewise-linear" form of the activation function. We thus write

$$0 < \mu(i) < 2\bar{\mu}(i)\,\frac{2}{f'(d(i)) + f'(\mathbf{u}_i\mathbf{w}_{i-1}) + \epsilon}, \tag{18}$$

where, for the sigmoid function, $f'(x) = \beta f(x)(1 - f(x))$. The positive number ϵ is introduced in order to avoid large step-sizes.

This approximation is however inconvenient in the cases when $\eta(i)$ happens to be close to zero, while $z(i)$ and $\mathbf{u}_i\mathbf{w}_{i-1}$ are reasonably far apart. To avoid a poor approximation in these cases, we may modify the above construction as follows: for improved convergence (i.e., with a disconnected feedback loop) we set

$$\mu_{opt}(i) = \frac{2\bar{\mu}(i)}{f'(d(i)) + f'(\mathbf{u}_i\mathbf{w}_{i-1}) + \epsilon} \tag{19}$$

if $\left(d(i) - \frac{1}{2}\right)\left(f(\mathbf{u}_i\mathbf{w}_{i-1}) - \frac{1}{2}\right) > 0$ or

$$\mu_{opt}(i) = \frac{\bar{\mu}(i)}{f'_{\max}} \tag{20}$$

otherwise. We refer to this construction as method B.

A third, and perhaps simpler method, is to first estimate $f(\eta(i))$ by the average of $f(d(i))$ and $f(\mathbf{u}_i\mathbf{w}_{i-1})$, i.e., $\hat{f}(\eta(i)) = 0.5[f(d(i)) + f(\mathbf{u}_i\mathbf{w}_{i-1}]$, and then set $f'(\eta(i)) \approx \beta\hat{f}(\eta(i))[1 - \hat{f}(\eta(i))]$. This leads to method C, with the choice

$$\mu_{opt}(i) = \frac{\bar{\mu}(i)}{\beta\left[\hat{f}(\eta(i))(1 - \hat{f}(\eta(i))\right] + \epsilon}. \tag{21}$$

9 Simulation Results

In all experiments, we have chosen a bipolar white random sequence with variance one as the input signal. We provide plots of learning curves for the error energy $|e_a(i)|^2$. The curves are averaged over 50 Monte Carlo runs in order to approximate $\mathrm{E}[|e_a(i)|^2]$. The weights to be identified were $\{1, 1, 1, 1, 1, 1, 1, 1, 1\}$. The first coefficient was used for the offset term while the other eight were driven by a bipolar input pattern. A neuron with these weights can be interpreted as one that finds the patterns with more than three +1.

The values of the inner signal z are from the set $\{-7,-5,-3,-1,1,3,5,7,9\}$. Since the 256 different input patterns consisted of the bipolar values $\{-1,+1\}$, we had $\|\mathbf{u}_i\|_2^2 = M$ and $\bar{\mu}(i) = 0.1111$ at every time instant i. We have chosen the sigmoid function (1) with $\beta = 0.4, 2, 4$. The first simulation is for $\beta = 0.4$, for which the sigmoid function operates in an almost linear range. The resulting learning curves are depicted in Figure 3. The learning curves are given in terms of the relative system mismatch defined as $S_{rel}(i) = E[\|\tilde{\mathbf{w}}_i\|_2^2]/\|\tilde{\mathbf{w}}_{-1}\|_2^2$.

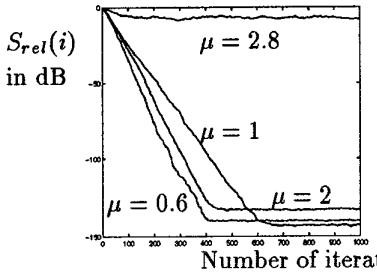

Figure 3: *Learning curves for Perceptron Learning Algorithm with* $\beta = 0.4$ *and* $\mu = 0.6, 1, 2, 2.8$.

As expected from (16), and since the sigmoid function operates essentially in the linear region, the fastest convergence speed occurs for $\mu = 10\bar{\mu}$ ($4/\beta = 10$), while instability occurs for values $\mu > 20\bar{\mu}$.

The next simulation shows learning curves for $\beta = 2$ (see Figure 4). With fixed step-sizes the fastest convergence was found at $\mu = 0.6$, while for $\mu = 1.2$ the algorithm was already unstable. The bound (16) for which the largest possible step-size is given by $\mu_l = 0.4444$ is now too conservative and the proposed modifications (A),(B) and (C) lead to much faster convergence. For all methods, the step-size was chosen to be optimal (with $T = 100$ and $\epsilon = 0.02$). Since method (C) always showed the same behaviour as (B) it is not depicted here. As the figure demonstrates, the first choice leads to excellent convergence, however, at the expense of calculating a logarithm at every time instant. The second choice, although not as perfect as the first one, still shows considerable improvement over the constant step-size choice.

For the third simulation $\beta = 4$. According to (16), convergence is expected for $\mu < 8\bar{\mu}/\beta = 2\bar{\mu} = 0.22222$. As Figure 5 shows, for μ smaller than this bound convergence occurs. However, this bound is rather conservative and fastest convergence occurs for larger step-size values, viz., $\mu \approx 0.4$. A learning curve for $\mu = 0.8$ still shows convergence but with some stopping effect. It seems noteworthy that even very large step-sizes can still lead to convergence, although the parameter estimates seem to diverge. This effect was not observed for small β and seems to arise from the fact that the system behaves highly nonlinearly. This effect could also be observed for $\beta = 8$, where it was even stronger.

Method (B), with the optimal choice for the step-size, was applied again and showed much faster convergence than any other choice of a constant step-size.

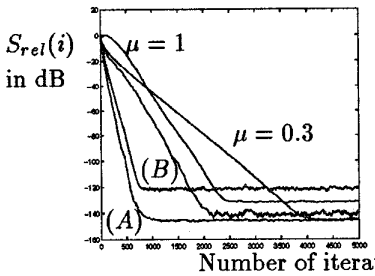

Figure 4: *Learning curves for Perceptron Learning Algorithm with* $\beta = 2$ *and* $\mu = 0.3, 0.6, 1$ *and* μ_{opt} *for methods (A) and (B).*

Instability occurred for approximately $2.2\mu_{opt}$.

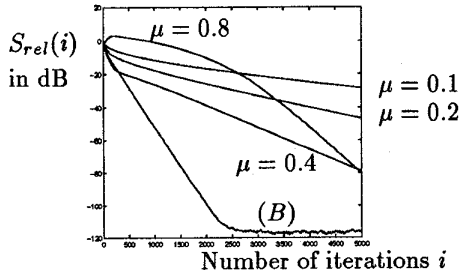

Figure 5: *Learning curves for Perceptron Learning Algorithm with* $\beta = 4$ *and* $\mu = 0.1, 0.2, 0.4, 0.8$ *and* μ_{opt} *from method (B).*

References

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation*, MacMillan Publishing Company, 1994.

[2] S. Hui, S.H. Zak, "The Widrow-Hoff algorithm for McCulloch-Pitts type neurons," *IEEE Trans. Neural Networks*, vol. 5, no. 6, pp. 924-929, Nov. 1994.

[3] D.R. Hush, B.G. Horne, "Progress in supervised neural networks," *IEEE Signal Processing Magazine*, vol. 10, no. 1, pp. 8-39, Jan. 1993.

[4] H. K. Khalil, *Nonlinear Systems*, MacMillan, 1992.

[5] R.P. Lippmann, "An introduction to computing with neural nets," *IEEE Acoustics, Speech and Signal Processing Magazine*, vol. 4, no. 2, pp. 4-22, April 1987.

[6] A.H. Sayed and M. Rupp, "A time-domain feedback analysis of adaptive gradient algorithms via the Small Gain Theorem", *Proc. SPIE Conference on Advanced Signal Processing*, vol. 2563, pp. 458–469, San Diego, CA, July 1995.

[7] J.J. Shynk and N.J. Bershad, "On the system identification convergence model for perceptron learning algorithms," *Proc. of Asilomar Conference on Signals, Systems, and Computers*, pp. 879-886, Oct. 1994.