Machine Learning Algorithms

Perceptron Learning Algorithm (PLA)

Markus Rupp

12.9.2020

Recall: Binary Classification

- Binary classification: y from {-1,1}

$$\underline{\hat{x}}_i^H \underline{\hat{c}} > 0; \quad if \quad y_i = +1$$

$$\underline{\hat{x}}_i^H \underline{\hat{c}} < 0; \quad if \quad y_i = -1$$

- We like to find an (iterative) algorithm that learns from data pairs and finds the hyperplane separating both categories

Sigmoid or tanh?

- For this we employ a nonlinear function f(x)
  - That maps the linear input (vector product) to one of the classes
  - And is sufficiently smooth, i.e., differentiable

$$\sigma\left(\underline{\hat{x}}_i^H \underline{c}\right) = \quad y_i + v_i$$

$$\sigma(x) = \frac{1}{1+e^{-\beta x}} \in [0,1]$$

$$2\sigma(x) - 1 = \frac{2}{1+e^{-\beta x}} - 1 = \frac{1-e^{-\beta x}}{1+e^{-\beta x}}$$

$$= \tanh\left(\frac{\beta x}{2}\right) \in [-1,1]$$

$$\frac{\partial \sigma(x)}{\partial x} = \beta \frac{e^{-\beta x}}{\left(1+e^{-\beta x}\right)^2} = \beta\sigma(x)\left(1-\sigma(x)\right)$$

$$\frac{\partial \tanh\left(\frac{\beta x}{2}\right)}{\partial x} = 2\beta\sigma(x)\left(1-\sigma(x)\right) = \frac{\beta}{2}\left(1-\tanh^2\left(\frac{\beta x}{2}\right)\right)$$

PLA

- In principle both are valid options
- Tanh is usually preferred due to symmetry
- For a function f(x) (sigmoid or tanh), we obtain the

- Perceptron Learning Algorithm (PLA)

$$\underline{\hat{\mathrm{w}}}_k = \underline{\hat{\mathrm{w}}}_{k-1} + \mu \underline{\mathrm{x}}_k^* \left(\mathrm{y}_k - f\left(\underline{\mathrm{x}}_k^T \underline{\hat{\mathrm{w}}}_{k-1} \right) \right); k = 1, 2, \ldots$$

PLA Derivation

- Let‘s derive this from first principles. We start with the cost function

$$c(\underline{w}) = \left(\mathrm{y}_k - f\left(\underline{\mathrm{x}}_k^T \underline{w} \right) \right)^2$$

- Its derivative is then given by

$$\frac{\partial}{\partial \underline{w}} c(\underline{w}) = -2\left(\mathrm{y}_k - f\left(\underline{\mathrm{x}}_k^T \underline{w} \right) \right) \; f'\left(\underline{\mathrm{x}}_k^T \underline{w} \right) \underline{\mathrm{x}}_k = \underbrace{-2e(\underline{w})\underline{\mathrm{x}}_k}_{\text{classical LS term}} f'\left(\underline{\mathrm{x}}_k^T \underline{w} \right)$$

PLA Derivation

- Using such gradient we find

$$\underline{\hat{\mathrm{w}}}_k = \underline{\hat{\mathrm{w}}}_{k-1} + \tilde{\mu} f'\left(\underline{\mathrm{x}}_k^T \underline{\hat{\mathrm{w}}}_{k-1}\right)\underline{\mathrm{x}}_k^*\left(\mathrm{y}_k - f\left(\underline{\mathrm{x}}_k^T \underline{\hat{\mathrm{w}}}_{k-1}\right)\right); k = 1,2,\ldots$$

- Since the sigmoid function ends up in a limited (positive) range of potential values, we consume the derivative into the new step-size μ and simplify the algorithmic update to obtain:

$$\underline{\hat{\mathrm{w}}}_k = \underline{\hat{\mathrm{w}}}_{k-1} + \mu \underline{\mathrm{x}}_k^*\left(\mathrm{y}_k - f\left(\underline{\mathrm{x}}_k^T \underline{\hat{\mathrm{w}}}_{k-1}\right)\right); k = 1,2,\ldots$$

Analysis

- Unfortunately, this small modification of the LMS algorithm leads to an unexpected hurdle in analysing its behaviour.
- Our classical form of analysis cannot be applied any longer and a new method needs to be applied.
- 1) show new method (based on energy relations) on LMS algorithm
- 2) apply it for PLA
- 3) apply it to recurrent neural nets (RNN)

Energy based relations

- As the entire analsis is based on the Cauchy Schwarz inequality, let‘s recall it….

Chauchy-Schwarz‘ Inequality

- **Theorem:** In an inner vector product space S in C^n with induced norm ||.|| the following holds:

$$\left|\left\langle \underline{x}, \underline{y} \right\rangle\right| \leq \left\|\underline{x}\right\|_2 \left\|\underline{y}\right\|_2$$

- The inequality holds with equal sign, if and only if, $\underline{x}=\alpha\underline{y}$.
- **Proof:** We start with the simple fact that

$$\left\|\underline{x} - \alpha\,\underline{y}\right\|_2 \geq 0$$

Chauchy-Schwarz‘ Inequality

- Furthermore, we have

$$0 \le \left\| \underline{x} - \alpha \underline{y} \right\|_2^2 = \left\| \underline{x} \right\|_2^2 - 2\,\mathrm{Re}\left\langle \underline{x}, \alpha \underline{y} \right\rangle + \left| \alpha \right|^2 \left\| \underline{y} \right\|_2^2$$

$$= \left\| \underline{x} \right\|_2^2 - \frac{\left| \left\langle \underline{x}, \underline{y} \right\rangle \right|^2}{\left\| \underline{y} \right\|_2^2} + \left\| \underline{y} \right\|_2^2 \left[\left(\alpha - \frac{\left\langle \underline{x}, \underline{y} \right\rangle}{\left\| \underline{y} \right\|_2^2} \right) \left(\alpha^* - \frac{\left\langle \underline{x}, \underline{y} \right\rangle^*}{\left\| \underline{y} \right\|_2^2} \right) \right]$$

- The minimum is obtained for: $\alpha = \dfrac{\left\langle \underline{x}, \underline{y} \right\rangle}{\left\| \underline{y} \right\|_2^2}$
- Thus:

$$0 \le \min_\alpha \left\| \underline{x} - \alpha \underline{y} \right\|_2^2 = \left\| \underline{x} \right\|_2^2 - \frac{\left| \left\langle \underline{x}, \underline{y} \right\rangle \right|^2}{\left\| \underline{y} \right\|_2^2} \qquad \rightarrow \left| \left\langle \underline{x}, \underline{y} \right\rangle \right| \le \left\| \underline{x} \right\|_2 \left\| \underline{y} \right\|_2$$

Chauchy-Schwarz‘ Inequality

- We are not done yet!

- For $\underline{x}=\alpha\underline{y}$ we obtain $||\underline{x}-\alpha\underline{y}||=0$.
- Due to its norm property $\|\underline{x}\| \geq 0$, this property $||\ \underline{x}-\alpha\underline{y}||=0$ can only be achieved for the argument to be zero.
- Thus, we obtain equality if and only if $\underline{x}=\alpha\underline{y}$.

Local Passivity Properties

- Now consider the term $\mu_k \left\| \underline{x}_k \right\|^2 \leq 1$
- Take an arbitrary estimate q.
- With Cauchy-Schwarz' inequality:

$$\frac{\left| \underline{x}_k^T \underline{w} - \underline{x}_k^T \underline{q} \right|^2}{\mu_k^{-1} \left\| \underline{w} - \underline{q} \right\|_2^2} \leq 1$$

- Alternatively:

$$\left| \underline{x}_k^T \underline{w} - \underline{x}_k^T \underline{q} \right|^2 - \mu_k^{-1} \left\| \underline{w} - \underline{q} \right\|_2^2 \leq 0$$

Local Passivity Properties

- Such relation is still valid, once we add something positive in the denominator

$$\frac{\left|\underline{x}_k^T \underline{w} - \underline{x}_k^T \underline{q}\right|^2}{\mu_k^{-1}\left\|\underline{w} - \underline{q}\right\|_2^2 + \left|v_k\right|^2} \leq 1$$

- Even for particular estimates, this is true:

$$\frac{\left|\underline{x}_k^T \underline{w} - \underline{x}_k^T \underline{\hat{w}}_{k-1}\right|^2}{\mu_k^{-1}\left\|\underline{w} - \underline{\hat{w}}_{k-1}\right\|_2^2 + \left|v_k\right|^2} \leq 1$$

Local Passivity Relations

- We are facing now the question, whether some algorithms with their specific estimates, like LMS or RLS, (or others) improve such relation even further by lowering the limit below one.

Local Passivity Properties

Theorem 5.1 (Local Passivity) For the adaptive gradient-type algorithm (LMS with variable step-size) for every time-instant the following properties are true:

$$a) \quad \frac{\mu_k^{-1}\left\|\underline{w}-\underline{\hat{w}}_k\right\|_2^2+\left|e_{a,k}\right|^2}{\mu_k^{-1}\left\|\underline{w}-\underline{\hat{w}}_{k-1}\right\|_2^2+\left|v_k\right|^2}\leq 1$$

$$b) \quad \frac{\left|e_{p,k}\right|^2+\left|e_{a,k}\right|^2}{\mu_k^{-1}\left\|\underline{w}-\underline{\hat{w}}_{k-1}\right\|_2^2+\left|v_k\right|^2}\leq 1$$

$$c) \quad \frac{\gamma_k\left\|\underline{w}-\underline{\hat{w}}_k\right\|_2^2+\left|e_{p,k}\right|^2}{\gamma_k\left\|\underline{w}-\underline{\hat{w}}_{k-1}\right\|_2^2+\left|v_k\right|^2}\leq 1$$

Local Passivity Properties

$$d) \quad \frac{\left|e_{a,k}\right|^2 + \left|e_{a,k+1}\right|^2}{\mu_k^{-1}\left\|\underline{w} - \underline{\hat{w}}_{k-1}\right\|_2^2 + \left|v_k\right|^2} \leq 1$$

The first three relations hold for $\mu_k \left\|\underline{x}_k\right\|_2^2 \leq 1$ **while for the last relation** $\mu_k \leq \min\left\{1/\left\|\underline{x}_k\right\|_2^2, 1/\left\|\underline{x}_{k+1}\right\|_2^2\right\}$

Also, $\gamma_k = \mu_k^{-1} - \left\|\underline{x}_k\right\|_2^2$

Local Passivity Relations

- **Proof:** Starting with the first relation, the update equation of the gradient type algorithm reads:

$$\underline{\widetilde{w}}_k = \underline{\widetilde{w}}_{k-1} - \mu_k \underline{x}_k^* \left[e_{a,k} + v_k \right]$$

Here, the disturbed error signal was split in the undisturbed part and the additive noise:

$$\widetilde{e}_{a,k} = e_{a,k} + v_k$$

Local Passivity Relations

- Computing the squared L_2 norm of both sides of the equation, we obtain:

$$\left\|\underline{\widetilde{w}}_k\right\|_2^2 = \left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2 + \mu_k^2 \left\|\underline{x}_k\right\|_2^2 \left|e_{a,k} + v_k\right|^2 - \mu_k \left[e_{a,k} + v_k\right] e_{a,k}^* - \mu_k \left[e_{a,k}^* + v_k^*\right] e_{a,k}$$

$$\left|e_{a,k} + v_k\right|^2 = \left|e_{a,k}\right|^2 + \left|v_k\right|^2 + v_k e_{a,k}^* + v_k^* e_{a,k}$$

- And thus:

$$\left\|\underline{\widetilde{w}}_k\right\|_2^2 = \left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2 + \mu_k^2 \left\|\underline{x}_k\right\|_2^2 \left|e_{a,k} + v_k\right|^2 - 2\mu_k \left|e_{a,k}\right|^2 - \mu_k \left[\left|e_{a,k} + v_k\right|^2 - \left|e_{a,k}\right|^2 - \left|v_k\right|^2\right]$$

Local Passivity Relations

- By reordering we finally obtain:

$$\left\|\underline{\widetilde{w}}_k\right\|_2^2 - \left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2 + \mu_k \left|e_{a,k}\right|^2 - \mu_k \left|v_k\right|^2$$

$$= \mu_k \left|e_{a,k} + v_k\right|^2 \left[\mu_k \left\|\underline{x}_k\right\|_2^2 - 1\right]$$

- As long as $\mu_k \left\|\underline{x}_k\right\|^2 \leq 1$ the right hand side remains negative, and thus

$$\left\|\underline{\widetilde{w}}_k\right\|_2^2 - \left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2 + \mu_k \left|e_{a,k}\right|^2 - \mu_k \left|v_k\right|^2 \leq 0$$

- Division gives us the desired result.

Robustness Analysis for LMS

- Need <u>global</u> relation from k=1,2,…,N

$$-\left\|\underline{\widetilde{w}}_k\right\|_2^2+\left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2+\mu_k\left|v_k\right|^2\geq\mu_k\left|e_{a,k}\right|^2$$

- Sum up and obtain:

$$\frac{\left\|\underline{\widetilde{w}}_N\right\|_2^2+\sum_{k=1}^{N}\mu_k\left|e_{a,k}\right|^2}{\left\|\underline{\widetilde{w}}_0\right\|_2^2+\sum_{k=1}^{N}\mu_k\left|v_k\right|^2}\leq 1$$

Robustness Analysis for LMS

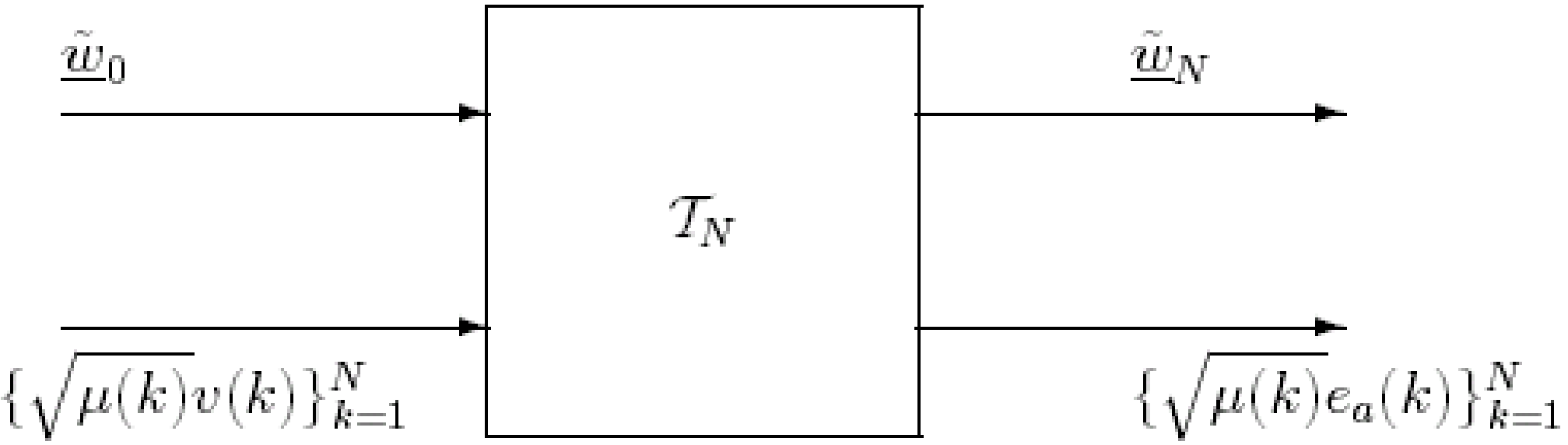
- Find passivity relation as well:

$$\begin{bmatrix} \sqrt{\mu_1}\, e_{a,1} \\ \vdots \\ \sqrt{\mu_N}\, e_{a,N} \\ \underline{\widetilde{w}}_N \end{bmatrix} = \underbrace{\begin{bmatrix} x & & & \\ x & x & & \\ x & x & x & \\ x & x & x & x \end{bmatrix}}_{\mathrm{T}_N} \begin{bmatrix} \underline{\widetilde{w}}_0 \\ \sqrt{\mu_1}\, v_1 \\ \vdots \\ \sqrt{\mu_1}\, v_1 \end{bmatrix}$$

- Induced matrix norm is bounded:

$$\left\| \mathrm{T}_N \right\|_{2,ind} \leq 1$$

Robustness Analysis for LMS



Minimax Optimality

- Is there a sequence so that the maximum one is really obtained?

$$\frac{\left\|\underline{\widetilde{w}}_N\right\|_2^2+\sum_{k=1}^{N}\mu_k\left|e_{a,k}\right|^2}{\left\|\underline{\widetilde{w}}_0\right\|_2^2+\sum_{k=1}^{N}\mu_k\left|v_k\right|^2}\overset{?}{=}1$$

Select special noise sequence

$$v_k=-e_{a,k}$$

Minimax Optimality

- However, in such case no update will occur, and the final estimate is identical to the first one, thus

$$\max_{\underline{\hat{w}}_0 \neq \underline{w}, v(.)} \frac{\left\| \underline{w} - \underline{\hat{w}}_N \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| e_{a,k} \right|^2}{\left\| \underline{w} - \underline{\hat{w}}_0 \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| v_k \right|^2} = 1$$

Minimax Optimality

- On the other hand, for <u>each</u> algorithm we have for v_k=-$e_{a,k}$ and :

$$\sum_{k=1}^{N} \mu_k \left| v_k \right|^2 = \sum_{k=1}^{N} \mu_k \left| e_{a,k} \right|^2 \leq \left\| \underline{w} - \underline{\hat{w}}_N \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| e_{a,k} \right|^2$$

- or equivalently (for $\underline{\hat{w}}_0 = \underline{w}$): $\underline{w} = \underline{\hat{w}}_0$

$$\max_{v(.)} \frac{\left\| \underline{w} - \underline{\hat{w}}_N \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| e_{a,k} \right|^2}{\left\| \underline{w} - \underline{\hat{w}}_0 \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| v_k \right|^2} \geq 1$$

Minimax Optimality

Theorem 5.2: (Minimax Property of the Gradient Type Algorithm) The gradient type algorithm solves the following Minimax problem if

$$\mu_k \left\| \underline{x}_k \right\|^2 \leq 1$$

$$\min_{\text{All algorithms}} \max_{\widetilde{w}_0, v(.)} \frac{\left\| \underline{w} - \underline{\hat{w}}_N \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| e_{a,k} \right|^2}{\left\| \underline{w} - \underline{\hat{w}}_0 \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| v_k \right|^2}$$

The optimal value is one.

Convergence

- **<u>Theorem 5.3</u> (Convergence Conditions)**

$$\textbf{If } \mu_k \left\| \underline{x}_k \right\|^2 \le 1, \left\| \underline{w}_0 \right\| < \infty, \textbf{ and } \sum_{k=1}^{\infty} \mu_k \left| v_k \right|^2 < \infty$$

then it follows that: $\sqrt{\mu_k}\, e_{a,k} \to 0.$

If, furthermore, $\sqrt{\mu_k}\, \underline{x}_k$ is persistent exciting, then we also have:

$$\underline{\hat{w}}_k \to \underline{w}$$

Convergence

- **Proof:** Since we have

$$\sum_{k=1}^{N} \mu_k \left| e_{a,k} \right|^2 \leq \left\| \underline{w} - \underline{\hat{w}}_0 \right\|_2^2 + \sum_{k=1}^{N} \mu_k \left| v_k \right|^2$$

- If initial error as well as noise energy is limited, then the a-priori error energy must be limited and thus must decay after some time to zero.
- Cauchy-Series!

Convergence

- Consider the update equation at time instant k:

$$\underline{\widetilde{w}}_k = \underline{\widetilde{w}}_{k-1} - \mu_k \underline{x}_k^* \left[e_{a,k} + v_k \right]$$

- Summing up over p time instants:

$$\underline{\widetilde{w}}_{k+p-1} = \underline{\widetilde{w}}_k - \sum_{l=1}^{p-1} \mu_{k+l} \widetilde{e}_{a,k+l} \underline{x}_{k+l}^*$$

- Consider a series of P>M vectors $\underline{x}_{k+p}$, p=1,…,P, all such vectors can be tested at one value of the parameter error vector:
-

$$x_{k+p}^T \underline{\widetilde{w}}_k = \underbrace{x_{k+p}^T \underline{\widetilde{w}}_{k+p-1}}_{e_{a,k+p}} - \sum_{l=1}^{p-1} \mu_{k+l} \widetilde{e}_{a,k+l} x_{k+p}^T \underline{x}_{k+l}^*$$

Convergence

- Since all $e_{a,k} \rightarrow 0$ and since also all $v_k \rightarrow 0$, we also have that

$$\tilde{e}_{a,k} \rightarrow 0$$

- Since the right hand side goes to zero, the left hand side must tend to zero as well. Since this is true for all vectors, we can write:

$$\begin{bmatrix} \underline{x}_{k+1}^T \\ \underline{x}_{k+2}^T \\ \vdots \\ \underline{x}_{k+P}^T \end{bmatrix} \underline{\tilde{w}}_k \rightarrow 0$$

Convergence

- From here we cannot conclude yet that $\underline{\widetilde{w}}_k \to \underline{0}$ since the vector could exist in the null space of $\underline{x}_k$.
- Invoking the persistent excitation condition as well, the matrix must have the full rank M, or equivalently:

$$\left[\underline{x}_{k+1}^{*}, \underline{x}_{k+2}^{*}, \cdots \underline{x}_{k+P}^{*}\right] \begin{bmatrix} \underline{x}_{k+1}^{T} \\ \underline{x}_{k+2}^{T} \\ \vdots \\ \underline{x}_{k+P}^{T} \end{bmatrix} \underline{\widetilde{w}}_k \to 0$$

Convergence

- Since the persistent excitation condition requires that

$$0 < \alpha I \leq \left[\underline{x}_{k+1}^{*}, \underline{x}_{k+2}^{*}, \ldots \underline{x}_{k+P}^{*}\right] \begin{bmatrix} \underline{x}_{k+1}^{T} \\ \underline{x}_{k+2}^{T} \\ \vdots \\ \underline{x}_{k+P}^{T} \end{bmatrix} \leq \beta I$$

- We can only conclude that

$$\underline{\widetilde{w}}_k \to \underline{0}$$

Small Gain Theorem

- Consider the Small Gain Theorem as an extension to what you know about stability of linear systems to those systems that are non-linear.
- Remember in linear system theory, the open loop gain is required to be smaller than one in order to have a closed loop gain bounded (BIBO stability).
- The Small Gain Theorem applies equally for non-linear systems as well as time-variant systems.
- Consider an input and output sequence as entries of a vector:

$$\underline{y}_N = H_N \underline{x}_N$$

Small Gain Theorem

- **Definition 1:** A mapping H is called l-stable if two positive constants β,γ exist, such that for all input sequences the following is true:

$$\left\| \underline{y}_N \right\| = \left\| H_N \underline{x}_N \right\| \leq \gamma \left\| \underline{x}_N \right\| + \beta$$

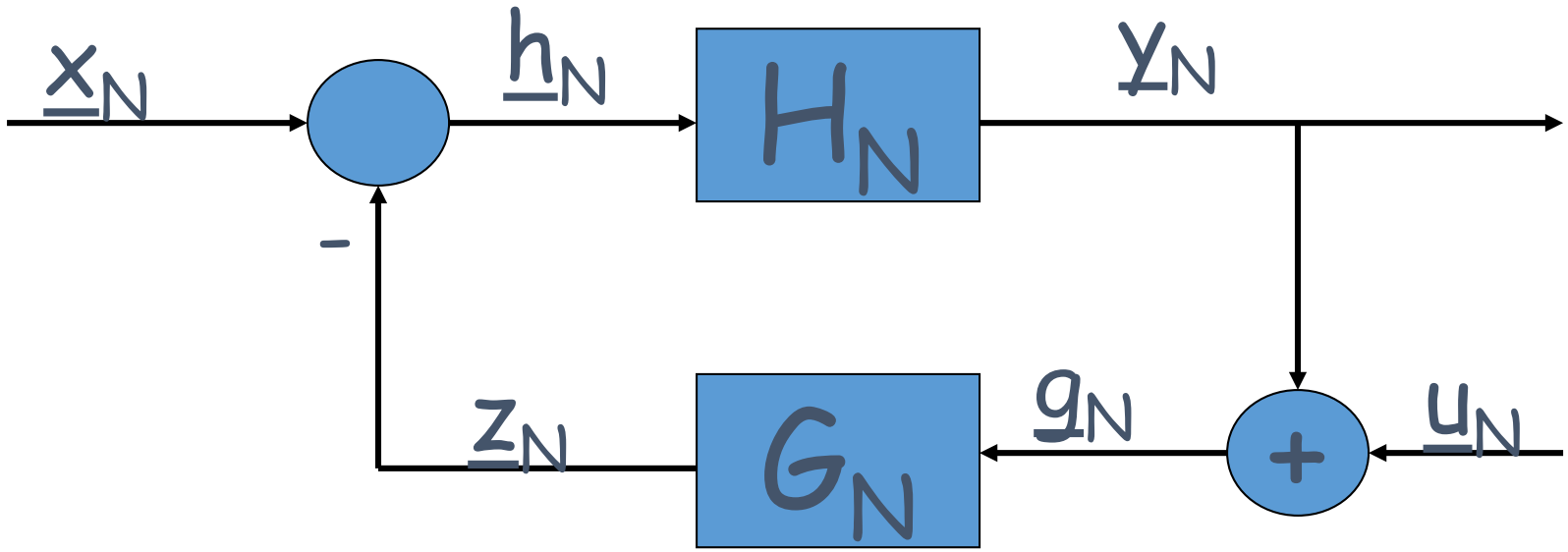
- **Definition 2:** The smallest positive value γ for which l-stability is guaranteed, is called the gain of H.
- **Remark:** BIBO (Bounded Input Bounded Output) stability is equivalent to l_{oo}- stability.

Small Gain Theorem

- Now consider a feedback system comprising of two systems G and H with corresponding gains γ_g and γ_h with the following signals:

$$\underline{y}_N = H_N \underline{h}_N = H_N \left[\underline{x}_N - \underline{z}_N \right]$$
$$\underline{z}_N = G_N \underline{g}_N = G_N \left[\underline{u}_N + \underline{y}_N \right]$$

Small Gain Theorem



Small Gain Theorem

- **Small Gain Theorem:** If the two gains γ_g and γ_h are such that

$$\gamma_g \gamma_h < 1$$

then the signals $\underline{g}_N$ and $\underline{h}_N$ of the feedback system are bounded by:

$$\left\| \underline{h}_N \right\| \leq \frac{1}{1-\gamma_g \gamma_h} \left[\left\| \underline{x}_N \right\| + \gamma_g \left\| \underline{u}_N \right\| + \beta_g + \gamma_g \beta_h \right]$$

$$\left\| \underline{g}_N \right\| \leq \frac{1}{1-\gamma_g \gamma_h} \left[\left\| \underline{u}_N \right\| + \gamma_h \left\| \underline{x}_N \right\| + \beta_h + \gamma_h \beta_g \right]$$

Small Gain Theorem

- **Proof:** Starting with the norm is given by:

$$\underline{h}_N = \underline{x}_N - \underline{z}_N; \quad \underline{g}_N = \underline{u}_N + \underline{y}_N$$

$$\begin{aligned}
\left\|\underline{h}_N\right\| &\le \left\|\underline{x}_N\right\| + \left\|G_N \underline{g}_N\right\| \\
&\le \left\|\underline{x}_N\right\| + \gamma_g \left\|\underline{g}_N\right\| + \beta_g \\
&\le \left\|\underline{x}_N\right\| + \gamma_g \left[\left\|\underline{u}_N\right\| + \gamma_h \left\|\underline{h}_N\right\| + \beta_h\right] + \beta_g \\
&= \gamma_g \gamma_h \left\|\underline{h}_N\right\| + \left\|\underline{x}_N\right\| + \gamma_g \left\|\underline{u}_N\right\| + \beta_g + \gamma_g \beta_h \\
&= \frac{1}{1-\gamma_g \gamma_h}\left[\left\|\underline{x}_N\right\| + \gamma_g \left\|\underline{u}_N\right\| + \beta_g + \gamma_g \beta_h\right]
\end{aligned}$$

- The derivation for $\underline{g}_N$ follows a similar path.

Robustness: Feedback Formulation

- For LMS we derived:

$$\left\|\underline{\widetilde{w}}_k\right\|_2^2 - \left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2 + \mu_k \left|e_{a,k}\right|^2 - \mu_k \left|v_k\right|^2$$

$$= \mu_k \left|e_{a,k} + v_k\right|^2 \left[\mu_k \left\|\underline{x}_k\right\|_2^2 - 1\right]$$

- Leading to three conditions, for

$$\mu_k \left\|\underline{x}_k\right\|_2^2 < 1 \rightarrow passivity$$

$$\mu_k \left\|\underline{x}_k\right\|_2^2 = 1 \rightarrow allpass,\ \mu_k = \overline{\mu}_k = \frac{1}{\left\|\underline{x}_k\right\|_2^2}$$

$$\mu_k \left\|\underline{x}_k\right\|_2^2 > 1 \rightarrow ?$$

Robustness: Feedback Formulation

- From a (temporary) local passivity relation for LMS

$$\frac{\left\|\underline{\widetilde{w}}_k\right\|_2^2+\mu_k\left|e_{a,k}\right|^2}{\left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2+\mu_k\left|v_k\right|^2}\begin{cases}\leq 1 & \textbf{if} \quad 0<\mu_k<\overline{\mu}_k\\ =1 & \textbf{if} \quad \mu_k=\overline{\mu}_k\\ >1 & \textbf{if} \quad \mu_k>\overline{\mu}_k\end{cases}$$

- We can derive a global relation

Feedback

- **Lemma 5.1:** For the gradient type algorithm the following is true at every time instant k (with $\overline{\mu}_k = 1/\left\|\underline{x}_k\right\|_2^2$):

$$\frac{\left\|\underline{\widetilde{w}}_k\right\|_2^2 + \mu_k \left|e_{a,k}\right|^2}{\left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2 + \mu_k \left|v_k\right|^2} \begin{cases} \leq 1 & \textbf{if} \quad 0 < \mu_k < \overline{\mu}_k \\ = 1 & \textbf{if} \quad \mu_k = \overline{\mu}_k \\ > 1 & \textbf{if} \quad \mu_k > \overline{\mu}_k \end{cases}$$

- **Proof:** The first has been proven already. The second can be directly obtained by setting $\mu_k = \overline{\mu}_k$. For the third one, consider again:

$$\left\|\underline{\widetilde{w}}_k\right\|_2^2 - \left\|\underline{\widetilde{w}}_{k-1}\right\|_2^2 + \mu_k \left|e_{a,k}\right|^2 - \mu_k \left|v_k\right|^2 = \mu_k \left|e_{a,k} + v_k\right|^2 \left[\underbrace{\mu_k \left\|\underline{x}_k\right\|_2^2}_{>1} - 1 \right]$$

Feedback

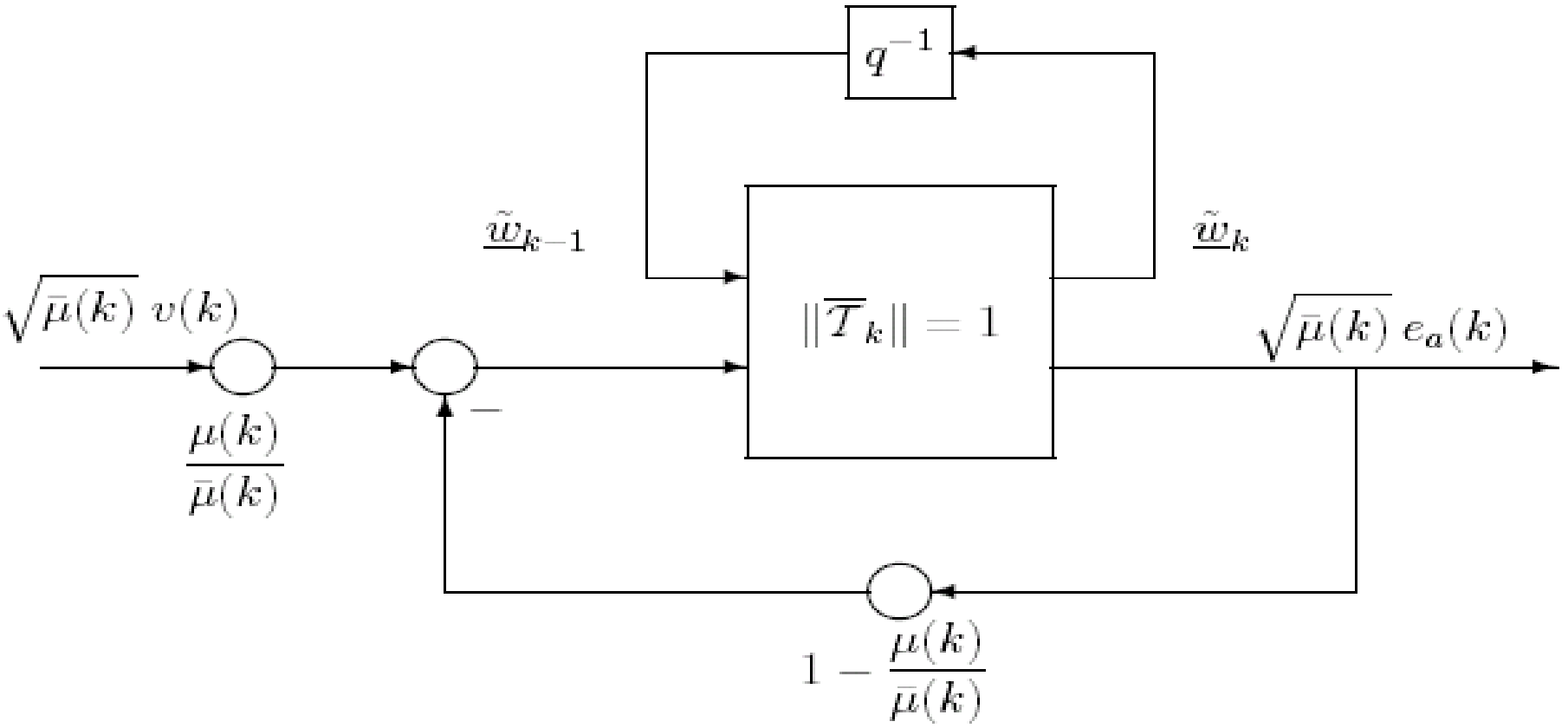
- Reformulation:

$$\underline{\hat{w}}_k = \underline{\hat{w}}_{k-1} + \mu_k \underline{x}_k^* \left[e_{a,k} + v_k \right]$$
$$= \underline{\hat{w}}_{k-1} + \overline{\mu}_k \underline{x}_k^* \left[e_{a,k} + \overline{v}_k \right]$$

- using

$$-\overline{v}_k = e_{a,k} - \frac{\mu_k}{\overline{\mu}_k} \left[e_{a,k} + v_k \right]$$
$$= \left[1 - \frac{\mu_k}{\overline{\mu}_k} \right] e_{a,k} - \frac{\mu_k}{\overline{\mu}_k} v_k$$

Feedback


q^{-1}
$\underline{\tilde{w}}_{k-1}$
$\underline{\tilde{w}}_k$
$\sqrt{\bar{\mu}(k)}\, v(k)$
$\|\overline{T}_k\| = 1$
$\sqrt{\bar{\mu}(k)}\, e_a(k)$
$-$
$\frac{\mu(k)}{\bar{\mu}(k)}$
$1 - \frac{\mu(k)}{\bar{\mu}(k)}$

Convergence of Feedback Scheme

- Consider again $$\sum_{k=1}^{N} \overline{\mu}_k \left| e_{a,k} \right|^2 \le \left\| \underline{w} - \underline{\hat{w}}_0 \right\|_2^2 + \sum_{k=1}^{N} \overline{\mu}_k \left| \overline{v}_k \right|^2$$

$$\overline{v}_k \overset{\Delta}{=} \left[\frac{\mu_k}{\overline{\mu}_k} \right] v_k + \left[\frac{\mu_k}{\overline{\mu}_k} - 1 \right] e_{a,k}$$

- Using short notations:

$$\delta(N) \overset{\Delta}{=} \max_{1 \le k \le N} \left| 1 - \frac{\mu(k)}{\overline{\mu}(k)} \right|$$

$$\gamma(N) \overset{\Delta}{=} \max_{1 \le k \le N} \frac{\mu(k)}{\overline{\mu}(k)}$$

Convergence of Feedback Scheme

$$\sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left|e_{a,k}\right|^2} \le \sqrt{\left\|\underline{\widetilde{w}}_0\right\|_2^2} + \sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left|\overline{v}_k\right|^2}$$

Use triangular inequality

$$\le \sqrt{\left\|\underline{\widetilde{w}}_0\right\|_2^2} + \sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left[\frac{\mu_k}{\overline{\mu}_k}\right]^2\left|v_k\right|^2} + \sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left[1-\frac{\mu_k}{\overline{\mu}_k}\right]^2\left|e_{a,k}\right|^2}$$

$$\le \left\|\underline{\widetilde{w}}_0\right\|_2 + \max\left|\frac{\mu_k}{\overline{\mu}_k}\right|\sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left|v_k\right|^2} + \max\left|1-\frac{\mu_k}{\overline{\mu}_k}\right|\sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left|e_{a,k}\right|^2}$$

$$= \left\|\underline{\widetilde{w}}_0\right\|_2 + \gamma_N\sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left|v_k\right|^2} + \delta_N\sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left|e_{a,k}\right|^2}$$

$$= \frac{1}{1-\delta_N}\left[\left\|\underline{\widetilde{w}}_0\right\|_2 + \gamma_N\sqrt{\sum_{k=1}^{N}\overline{\mu}_k\left|v_k\right|^2}\right]$$

Convergence of Feedback Scheme

- **<u>Theorem 5.4</u> (Extended Convergence of the Gradient Type Algorithm)**

$$\textbf{If } \mu_k \left\| \underline{x}_k \right\|^2 < 2, \left\| \underline{\tilde{w}}_0 \right\| < \infty, \textbf{ and } \sum_{k=1}^{\infty} \overline{\mu}_k \left| v_k \right|^2 < \infty$$

then it follows that: $\sqrt{\overline{\mu}_k}\, e_{a,k} \to 0.$

If, furthermore, $\sqrt{\mu_k}\, \underline{x}_k$ is persistent exciting, then we also have:

$$\underline{\hat{w}}_k \to \underline{w}$$

Convergence of Feedback Scheme

- Also consider the energy flow in such system. Since the forward path has no loss of energy, all energy going in must come out. The energy part in the parameter error vector is fed back. Only here in the feedback path energy can be lost. The more energy is lost, the faster is the adaptation. Fastest adaptation is thus obtained for

$$\mu_k = \overline{\mu}_k$$

Nonlinear, Memoryless Filter in the Estimation Path

- Consider now $f[y_k]$ instead of y_k

- Perceptron Learning Algorithm

Perceptron Learning Algorithm

- Consider two sets S_1 and S_2 of M-dimensional real-valued vectors $\underline{x}_k$

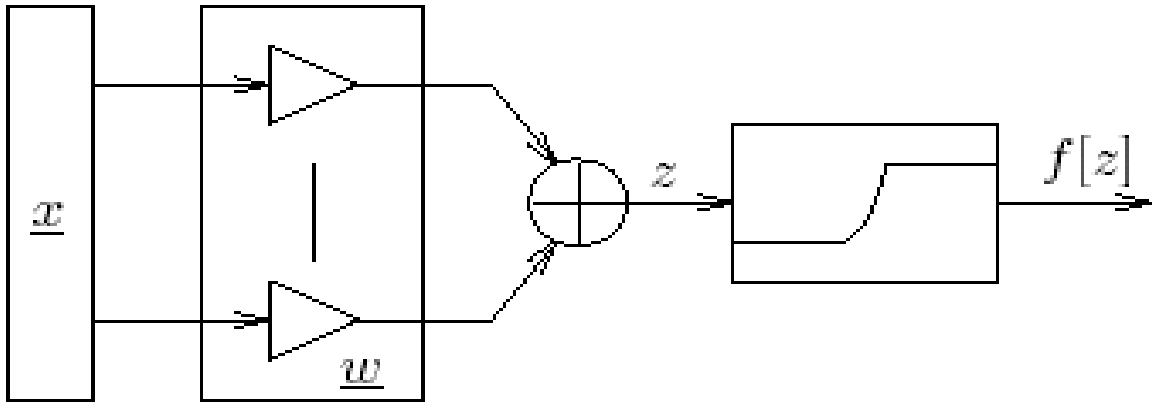
$$S_1 = \left\{ \underline{x}_k \in R^M \mid \underline{x}_k \text{ has property A} \right\}$$

$$S_2 = \left\{ \underline{x}_k \in R^M \mid \underline{x}_k \text{ has property } \overline{\text{A}} \right\}$$

- If both sets are linearly separable, then they describe two classes.

Perceptron Learning Algorithm

- Model of linear perceptron



Perceptron Learning Algorithm

- Activity function is in general a sigmoid-function

- In particular:

$$f_\beta[z] = \frac{1}{1+e^{-\beta z}}; \quad \beta > 0$$

$$f_\infty[z] = \frac{1+\mathrm{sgn}[z]}{2} = \begin{cases} 0 & ;\textbf{if } z < 0 \\ 1/2 & ;\textbf{if } z = 0 \\ 1 & ;\textbf{if } z > 0 \end{cases}$$

Perceptron Learning Algorithm

- Consider input vectors $\underline{x}_k$ with their corresponding output values y_k

$$y_k = f[\underline{x}_k^T \underline{w}]$$

- In supervised learning, the data pairs $\{\underline{x}_k, y_k\}$ are offered to the synapse in order to find $\underline{w}$.

Perceptron Learning Algorithm

- The most well-known algorithm with such property is the PLA

$$\underline{\hat{w}}_k = \underline{\hat{w}}_{k-1} + \mu \underline{x}_k \left(y_k - f\left[\underline{x}_k^T \underline{\hat{w}}_{k-1} \right] \right)$$

- Consider again the general form including additive noise

$$d_k = f\left[\underline{x}_k^T \underline{w} \right] + v_k = y_k + v_k$$

- This leads to:

$$\underline{\hat{w}}_k = \underline{\hat{w}}_{k-1} + \mu_k \underline{x}_k \left(d_k - f\left[\underline{x}_k^T \underline{\hat{w}}_{k-1} \right] \right)$$

Perceptron Learning Algorithm

- Note that this update

$$\underline{\hat{w}}_k = \underline{\hat{w}}_{k-1} + \mu \underline{x}_k \left(y_k - f\left[\underline{x}_k^T \underline{\hat{w}}_{k-1} \right] \right)$$

- is a somewhat simplified form of

$$\underline{\hat{w}}_k = \underline{\hat{w}}_{k-1} + \tilde{\mu} \underline{x}_k \left(y_k - f\left[\underline{x}_k^T \underline{\hat{w}}_{k-1} \right] \right) f'\left[\underline{x}_k^T \underline{\hat{w}}_{k-1} \right]$$

- Where we incorporated the derivative into the stepsize

Perceptron Learning Algorithm

- Because of the mean value theorem we have

$$f\left[\underline{x}_k^T \underline{w}\right] - f\left[\underline{x}_k^T \underline{\hat{w}}_{k-1}\right] = f'\left[\eta_k\right] e_{a,k}$$

- For
- With this, the PLA can be treated as the previous algorithms.

$$\underline{x}_k^T \underline{w} < \eta_k < \underline{x}_k^T \underline{\hat{w}}_{k-1}$$

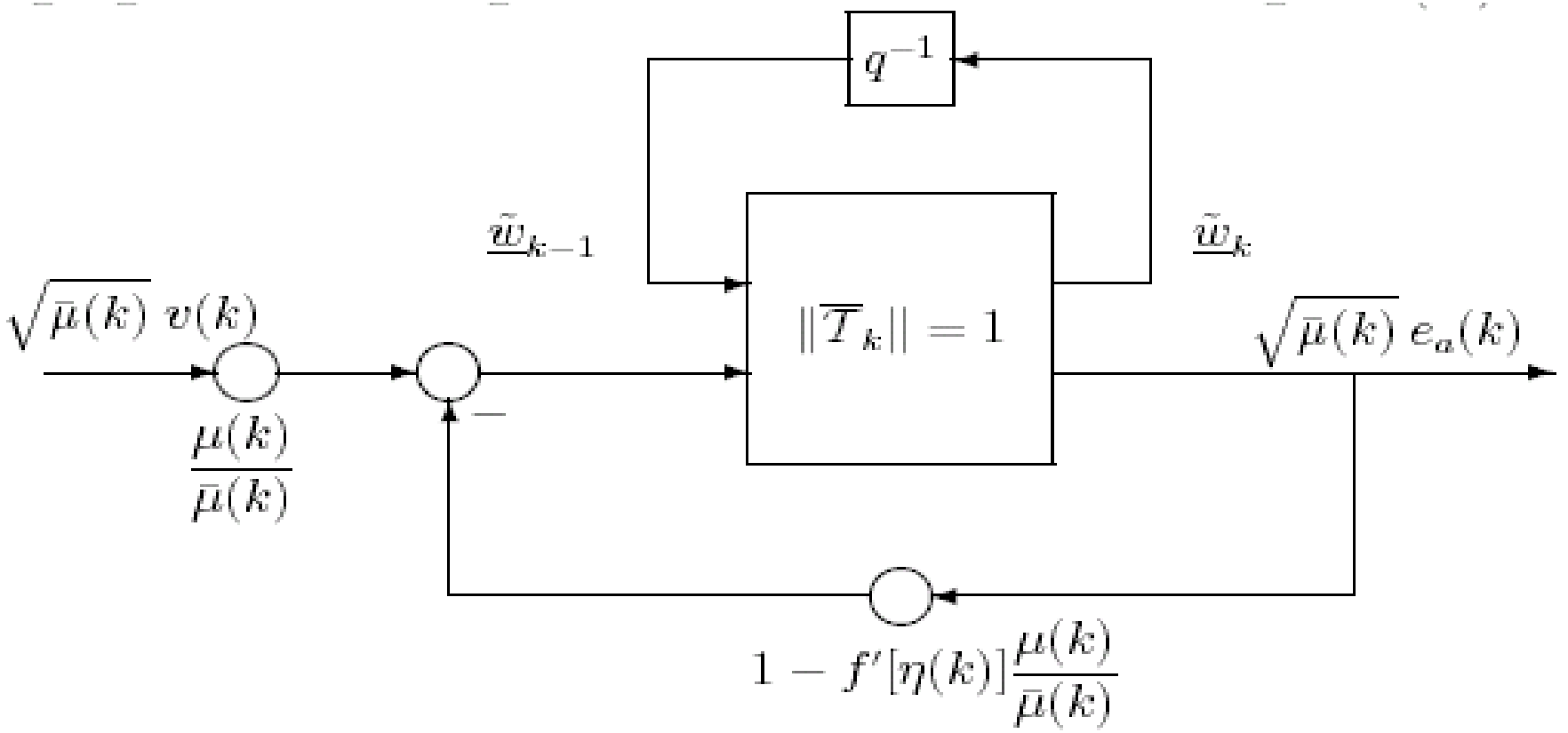
Perceptron Learning Algorithm

- We find again the stability condition

$$\delta_N = \max_{1 \le k \le N} \left| 1 - f'[\eta_k] \frac{\mu_k}{\overline{\mu}_k} \right|$$

- Together with a feedback structure

Perceptron Learning Algorithm



PLA

- Different to the LMS algorithm, the PLA includes a derivative of the nonlinear function in its feedback path
- As we (mostly) do not know in advance at which point we have to take the derivative, we take its largest value as upper bound
- Sigmoid function: max derivative is ¼ β
- Tanh function: max derivative is ½ β

- Leads to conservative results but ensures stability!

PLA

- Recall that we have incorporated the derivative into the step-size
- If we would not have done so, the derivative would appear twice in the feedback loop
  - However, not exactly the same term as one goes with

$$\underline{x}_k^T \, \underline{\hat{w}}_{k-1}$$

  - While the other goes with

$$\underline{x}_k^T \, \underline{w} < \eta_k < \underline{x}_k^T \, \underline{\hat{w}}_{k-1}$$

  - We will see in the backpropagation algorithm, the update will change…

Linear and Non-Linear Filter in Reference Path

- Recurrent Neuronal Networks (RNN) are very suitable for classification
- The structure proposed by Narendra and Parthasarathy is often treated in literature since it is suitable for calculation.

Linear and Non-Linear Filter in Reference Path

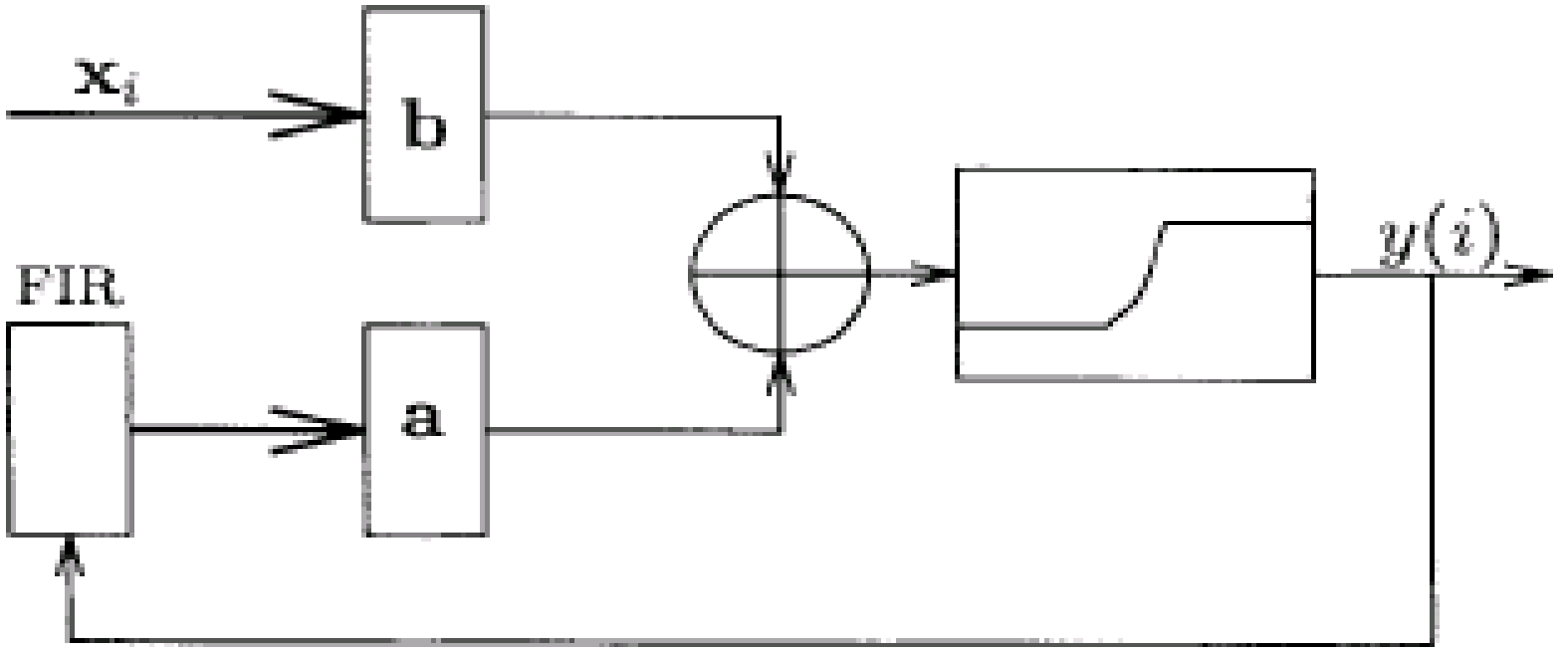


Fig. 6. Narendra and Parthasarathy's dynamic network.

Linear and Non-Linear Filter in Reference Path

- The update equation can be written as:

- With the values $$\underline{\hat{w}}_k = \underline{\hat{w}}_{k-1} + \mu_k \underline{u}_k \left[\frac{1}{f'^{-1}\left[\eta_k\right] - A\left(q^{-1}\right)} \left[e_{a,k}\right] + v_k \right]$$

- We find:

$$\lambda = \min_k \left\{ f'^{-1}\left[\eta_k\right] \right\}, \quad \zeta^{-1} = \max_k \left\{ \overline{\mu}_k^{-1} \right\}$$

$$\mathrm{Real}\left\{1 - A\left(e^{j\Omega}\right)\right\} < \delta, \quad \Omega \in \left[0, 2\pi\right]$$

$$\mu < 2\zeta\left(\lambda + \delta - 1\right)$$

Linear and Non-Linear Filter in Reference Path

- There is a direct relation to the slope of the derivative of the sigmoid function:

$$\beta < \frac{8}{\frac{\mu}{\zeta} + 2(1-\delta)}$$

- Note that although nonlinearity and a linear filter occur in the reference path, it is equivalently a system with nonlinear filter in the error path!

Literature

- M. Rupp, A.H. Sayed:
  "*Supervised Learning of Perceptron and Output Feedback Dynamic Networks: A Feedback Analysis via the Small Gain Theorem*";
  IEEE Transactions on Neural Networks, **Vol. 8** (1997), no. 3; S. 612 - 622.