In *Theoretical Advances in Neural Computation and Learning,* V. Roychowdhury, K. Siu and A. Orlitsky, eds., Kluwer Academic Publishers, Ch. 12, pp. 425–447, Norwell, MA, 1994.

LMS and Backpropagation are Minimax Filters *

Babak Hassibi† Ali H. Sayed and Thomas Kailath

Abstract

We give a minimax interpretation of instantaneous-gradient-based learning algorithms such as LMS and backpropagation. When the underlying model is linear, we show that the LMS algorithm minimizes the worst case ratio of predicted error energy to disturbance energy. When the model is nonlinear models, that arises in the context of neural networks, we show that the backpropagation algorithm performs this minimization in a *local* sense. These results provide theoretical justification of the widely observed excellent robustness properties for the LMS and backpropagation algorithms.

*This work was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command under Contract AFOSR91-0060 and by the Army Research Office under contract DAAL03-89-K-0109.

†Information Systems Laboratory, Stanford University, Stanford CA 94305.

1 Introduction

An important problem that arises in many applications is the following adaptive problem: given a sequence of $n \times 1$ input column vectors $\{h_i\}$, and a corresponding sequence of desired scalar responses $\{d_i\}$, find an estimate of an $n \times 1$ column vector of weights w such that the sum of squared errors, $\sum_{i=0}^{N} |d_i - h_i^T w|^2$, is minimized. The $\{h_i, d_i\}$ are most often presented sequentially, and one is therefore required to find an adaptive scheme that recursively updates the estimate of w. The least-mean-squares (LMS) algorithm was originally conceived as an approximate solution to the above adaptive problem, that recursively updates the estimates of the weight vector along the direction of the *instantaneous gradient* of the sum squared error [1]. The introduction of the LMS adaptive filter in 1960 came as a significant development for a broad range of engineering applications since the LMS adaptive linear-estimation procedure requires essentially no advance knowledge of the signal statisitics. The LMS, however, has been long thought to be an approximate minimizing solution to the above squared error criterion, and a rigorous minimization criterion has been missing.

Exact recursive-least-squares (RLS) algorithms have also been developed (see, *e.g.* [2]). These algorithms have better covergence properties, but are computationally more complex and, in the presence of model uncertainties and lack of statistical information, exhibit poorer robust behaviour than the simple LMS. For example, it has been observed that the LMS has better tracking capabilities than the RLS algorithm in the presence of nonstationary inputs [2].

The nonlinear counterpart of the afore mentioned problem is one in which we are given a sequence of nonlinear functions $\{h_i(.)\}$ (*e.g. sigmoids*), and a corresponding sequence of desired responses $\{d_i\}$, and are required to recursively construct an estimate of the weight vector w such that $\sum_{i=0}^{N} |d_i - h_i(w)|^2$ is minimized. To date, exact solutions to this minimization problem for general nonlinear functions $h_i(.)$ do not exist, and the celebrated *backpropagation algorithm* is an approximate recursive solution [3, 4, 5] that updates the weight vectors along the direction of the instantaneous gradient. In this sense the backpropagation algorithm is an extension of the LMS algorithm to the nonlinear setting encountered in neural networks. Backpropagation has aslo proven to be a very robust algorithm in practice, and is currently the most widely used algorithm for training adaptive neural networks.

In this paper we provide a minimax interpretation of such instantaneous-gradient-based learning algorithms. In particular, we show that the LMS algorithm is an H^∞ optimal filter, where the H^∞ norm has been recently introduced as a robust criterion for problems in estimation and control [6]. In other words, LMS minimizes the worst case ratio of prediction energy to disturbance energy. Thus the LMS algorithm, which has long been regarded as an approximate least-squares solution, is in fact a minimizer of the H^∞ norm, and not of the H^2 norm. We further extend this result to the nonlinear setting that often arises in the study of neural networks, and show that the backpropagation algorithm is *locally* H^∞ optimal. These statements will be made precise in the next few sections.

Our results yield a new interpretation of instantaneous gradient based adaptive algorithms, and readily provide theoretical justification for the widely observed excellent robustness and tracking properties of the LMS and backpropagation algorithms. These algorithms guarantee the smallest estimation error energy over all possible disturbances of fixed energy, and are thus over conservative, which reflects in better robust behaviour to disturbance variation. Moreover, we are lead to an interesting connection between these learning algorithms and the emerging field of H^∞ estimation [7].

In this paper, we attempt to introduce the main concepts, motivate the results, and discuss the various implications. We essentially outline the proofs, and the

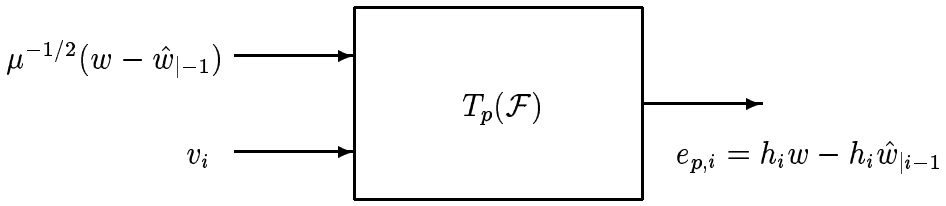

Figure 1: Transfer operator form disturbances to prediction errors.

reader is referred to [8] for more details and for connections to H^∞ estimation theory.

2 Linear Minimax Adaptive Filtering

We begin with the definition of the H^∞ norm of a transfer operator. As will presently become apparent, the motivation for introducing the H^∞ norm is to capture the worst case behaviour of a system.

Let h_2 denote the vector space of square-summable complex-valued causal sequences $\{f_k,\ 0 \leq k < \infty\}$, viz.,

$$h_2 = \{\text{set of sequences } \{f_k\} \text{ such that } \sum_{k=0}^{\infty} f_k^* f_k < \infty\}$$

with inner product $< \{f_k\}, \{g_k\} > = \sum_{k=0}^{\infty} f_k^* g_k$, where $*$ denotes complex conjugation. Let T be a transfer operator that maps a causal input sequence $\{u_i\}$ to a causal output sequence $\{y_i\}$. Then the H^∞ norm of T is given by

$$\|T\|_\infty = \sup_{u \in h_2, u \neq 0} \frac{\|y\|_2}{\|u\|_2}$$

where the notation $\|u\|_2$ denotes the h_2-norm of the causal sequence $\{u_k\}$, viz.,

$$\|u\|_2^2 = \sum_{k=0}^{\infty} u_k^* u_k.$$

The H^∞ norm may be thus regarded as the maximimum *energy gain* from the input u to the output y.

2.1 Formulation of the Problem

Suppose we observe an output sequence $\{d_i\}$ that obeys the following model:

$$d_i = h_i^T w + v_i \tag{1}$$

where $h_i^T = \begin{bmatrix} h_{i1} & h_{i2} & \dots & h_{in} \end{bmatrix}$ is a known input vector, w is an unknown weight vector, and $\{v_i\}$ is an unknown disturbance, which may also include modeling errors. We shall not make any assumptions on the noise sequence $\{v_i\}$, such as stationarity, whiteness, etc.

Let $w_i = \mathcal{F}(d_0, d_i, \dots, d_i)$ denote the estimate of the weight vector w given the observations $\{d_j\}$ from time 0 up to and including time i. The objective is to determine the functional $\mathcal{F}$, and consequently the estimate w_i, so as to minimize a certain norm defined in terms of the prediction error

$$e_i = h_i^T w - h_i^T w_{i-1}$$

which is the difference between the true (uncorrupted) output $h_i^T w$ and the predicted output $h_i^T w_{i-1}$. Let T denote the transfer operator that maps the unknowns $\{\mu^{-\frac{1}{2}} w - w_{-1}, v_i\}$, where w_{-1} denotes an initial guess of w and μ is a positive constant, to the prediction error e_i. See Figure 1. The H^∞ estimation problem can now be formulated as follows.

Problem 1 (Minimax Adaptive Problem) *Find an H^∞-optimal estimation strategy $w_i = \mathcal{F}(d_0, d_1, \ldots, d_i)$ that minimizes $\|T\|_\infty$, and obtain the resulting*

$$\gamma_0^2 = \inf_{\mathcal{F}} \|T\|_\infty^2 = \inf_{\mathcal{F}} \sup_{w,v \in h_2} \frac{\|e\|_2^2}{\mu^{-1}|w - w_{-1}|^2 + \|v\|_2^2} \tag{2}$$

where $|w - w_{-1}|^2 = (w - w_{-1})^T (w - w_{-1})$, and μ is a positive constant that reflects apriori knowledge as to how close w is to the initial guess w_{-1}. ■

Note that the infimum in (2) is taken over all *causal* estimators $\mathcal{F}$. This is of significance since the solution of the H^∞ estimation problem depends on the structure of the information available to $\mathcal{F}$. The above problem formulation shows that H^∞ optimal estimators guarantee the smallest prediction error energy over all possible disturbances of fixed energy. H^∞ estimators are thus over conservative, which reflects in a more robust behaviour to disturbance variation. Expression (2) clearly indicates the minimax nature of H^∞ optimal estimation.

At this point we need one more definition.

Definition 1 (Exciting Inputs) *The input vectors h_i are called exciting if, and only if,*

$$\lim_{N \to \infty} \sum_{i=0}^{N} h_i^T h_i = \infty$$

■

2.2 Main Result

Theorem 1 (LMS Algorithm) *Consider the model (1), and suppose we wish to minimize the H^∞ norm of the transfer operator from the unknowns $w - w_{-1}$ and v_i to the prediction error e_i. If the input vectors are exciting and*

$$0 < \mu < \inf_i \frac{1}{h_i^T h_i} \tag{3}$$

then the minimum H^∞ norm is $\gamma_{opt} = 1$. In this case an optimal H^∞ estimator is given by the LMS algorithm with learning rate μ, viz.

$$w_i = w_{i-1} + \mu h_i (d_i - h_i^T w_{i-1}) \qquad , \qquad w_{-1} \tag{4}$$

■

In other words, the result states that the LMS algorithm is an $H^\infty-$optimal filter. Moreover, Theorem 1 also gives an upper bound on the learning rate μ that ensures the H^∞ optimality of LMS. This is in accordance with the well-known fact that LMS behaves poorly if the learning rate is too large. It is also interesting to compare (3) with the bound obtained in [9], *viz.*,

$$\mu < \frac{1}{E[h_i^T h_i]}$$

where the $\{h_i\}$ are assumed to be random variables, and the $\{v_i\}$ are assumed to be independent white noise.

Intuitively it is not hard to convince oneself that γ_{opt} cannot be less than one. To this end suppose that the estimator has chosen some initial guess w_{-1}. Then one may conceive of a disturbance that yields an observation that coincides with the output expected from w_{-1}, i.e.

$$h_i^T w_{-1} = h_i^T w + v_i = d_i$$

In this case one expects that the estimator will not change its estimate of w, so that $w_i = w_{-1}$ for all i. Thus the prediction error will be

$$e_i = h_i^T w - h_i^T w_{i-1} = h_i^T w - h_i^T w_{-1} = -v_{i-1}$$

and the ratio in (2) can be made arbitrarily close to one.

The surprising fact though is that γ_{opt} is one and that the LMS algorithm achieves it. What this means is that LMS guarantees that the energy of the prediction error will never exceed the energy of the disturbances. This is not true for other estimators. For example, in the case of the recursive least-squares (RLS) algorithm, one can come up with a disturbance of arbitrarily small energy that will yield a prediction error of large energy.

Proof of Theorem 1: We give here an outline of the proof. For more details and for connections with H^∞ estimation see [8].

We first study the problem of bounding $\| T \|_\infty$ by a constant $\gamma > 0$. Let T_i denote the transfer operator that maps the disturbances $\{\mu^{\frac{1}{2}}(w - w_{-1}), \{v_j\}_{j=0}^{i-1}\}$ to the prediction erros $\{\{e_j = h_j^T w - h_j^T w_{j-1}\}_{j=0}^{i}\}$. In order to guarantee $\| T \|_\infty < \gamma$ we shall ensure $\| T_i \|_\infty < \gamma$ for all $i \geq 0$.

¿From the definition of the H^∞ norm of a transfer operator, this implies that for all $w \neq w_{-1}$ and all nonzero $v \in h_2$ we must find estimates w_j such that

$$\frac{\sum_{j=0}^{i} |e_j|^2}{\mu^{-1}|w - w_{-1}|^2 + \sum_{j=0}^{i-1} |v_j|^2} < \gamma^2$$

Since the denominator in the above inequality is nonzero we can write

$$\mu^{-1}|w - w_{-1}|^2 + \sum_{j=0}^{i-1} |v_j|^2 - \gamma^{-2} \sum_{j=0}^{i} |e_j|^2 > 0$$

or equivalently,

$$J_i = \mu^{-1}|w - w_{-1}|^2 + \sum_{j=0}^{i-1} |d_j - h_j^T w|^2 - \gamma^{-2} \sum_{j=0}^{i} |h_j^T w_{j-1} - h_j^T w|^2 > 0$$

Thus we must find estimates w_j such that J_i is positive for all $w \neq w_{-1}$. Since J_i is a quadratic form in w, we must have a minimum over w, otherwise w can be chosen to make J_i arbitrarily negative. For this to happen we need

$$\frac{\partial^2 J_i}{\partial w^2} = \mu^{-1} I + \sum_{j=0}^{i-1} h_j h_j^T - \gamma^{-2} \sum_{j=0}^{i} h_j h_j^T > 0$$

Therefore

$$\mu^{-1} I + (1 - \gamma^{-2}) \sum_{j=0}^{i-1} h_j h_j^T - \gamma^{-2} h_i h_i^T > 0 \qquad (5)$$

Suppose $\gamma < 1$ so that $1 - \gamma^{-2} < 0$. Since the $\{h_j\}$ are exciting, we conclude that for some k, and for large enough i, we must have

$$\sum_{j=0}^{i-1} |h_{jk}|^2 > \frac{\mu^{-1}}{\gamma^{-2} - 1}$$

This implies that the k^{th} diagonal entry of the matrix in (5) is negative, viz.,

$$\mu^{-1} + (1 - \gamma^{-2}) \sum_{j=0}^{i-1} |h_{jk})|^2 - \gamma^{-2} |h_{ik}|^2 < 0$$

Consequently, $\mu^{-1} I + (1 - \gamma^{-2}) \sum_{j=0}^{i-1} h_j h_j^T - \gamma^{-2} h_i h_i^T$ cannot be positive-definite. Therefore, $\gamma_{opt} \geq 1$. Suppose now that $\gamma = 1$. Then (5) reduces to

$$\mu^{-1} I - \gamma^2 h_j h_j^T > 0 \tag{6}$$

It is straightforward to see that the the eigenvalues of the matrix in (6) are

$$\{\mu^{-1}, \mu^{-1}, ..., \mu^{-1}, \mu^{-1} - h_i h_i^T\}$$

Thus (6) is satisfied if, and only if, (3) is satisfied.

Now that we have guaranteed that J_i has a minimum over w, we must show that the estimate given by LMS renders J_i positive for $\gamma = 1$. We shall do so by induction. For $i = 0$ we have:

$$J_0 = \mu^{-1} |w - w_{-1}|^2 - |h_0^T w - h_0^T w_{-1}|^2 = (w - w_{-1})^T \left[\mu^{-1} I - h_0 h_0^T\right] (w - w_{-1}) > 0$$

since $w \neq w_{-1}$ and $(\mu^{-1} - h_0 h_0^T)$ is positive definite by (3). For $i = 1$we have:

$$\begin{aligned}
J_1 &= \mu^{-1} |w - w_{-1}|^2 - |h_0^T w - h_0^T w_{-1}|^2 + |d_0 - h_0^T w|^2 \\
&\quad -|h_1^T w - h_1^T w_0|^2 \\
&= \mu^{-1} |w - w_{-1}|^2 - |h_0^T w - h_0^T w_{-1}|^2 + |d_0 - h_0^T w|^2 \\
&\quad -|h_1^T \left(w - w_{-1} - \mu h_0 (d_0 - h_0^T w_{-1})\right)|^2 \\
&= \mu^{-1} |w - w_{-1}|^2 - |h_0^T (w - w_{-1})|^2 + |(d_0 - h_0^T w_{-1}) - h_0^T (w - w_{-1})|^2 \\
&\quad -|h_1^T (w - w_{-1}) - \mu h_1^T h_0 (d_0 - h_0^T w_{-1})|^2 \\
&= \begin{bmatrix} w - w_{-1} \\ d_0 - h_0^T w_{-1} \end{bmatrix}^T \\
&\quad \begin{bmatrix} \mu^{-1} I - h_0 h_0^T + h_0 h_0^T - h_1 h_1^T & -h_0 + \mu h_1 h_1^T h_0 \\ -h_0^T + \mu h_0^T h_1 h_1^T & 1 - \mu^2 h_0^T h_1 h_1^T h_0 \end{bmatrix} \begin{bmatrix} w - w_{-1} \\ d_0 - h_0^T w_{-1} \end{bmatrix} \\
&= \begin{bmatrix} w - w_{-1} \\ d_0 - h_0^T w_{-1} \end{bmatrix}^T \\
&\quad \begin{bmatrix} \mu^{-1} I - h_1 h_1^T & -\mu(\mu^{-1} I - h_1 h_1^T) h_0 \\ -\mu h_0^T (\mu^{-1} I - h_1 h_1^T) & 1 - \mu^2 h_0^T h_1 h_1^T h_0 \end{bmatrix} \begin{bmatrix} w - w_{-1} \\ d_0 - h_0^T w_{-1} \end{bmatrix} > 0
\end{aligned} \tag{7}$$

Where, for the last inequality, we have used the fact that the matrix appearing in (7) is positive definite. To show this we note that the $(1,1)$ element $\mu^{-1} I - h_1 h_1^T$ is positive definite, and the Schur complement [10]

$$1 - \mu^2 h_0^T h_1 h_1^T h_0 - \mu h_0^T (\mu^{-1} I - h_1 h_1^T)(\mu^{-1} I - h_1 h_1^T)^{-1} (\mu^{-1} I - h_1 h_1^T) h_0 \mu =$$
$$1 - \mu^2 h_0^T h_1 h_1^T h_0 - \mu h_0^T (\mu^{-1} I - h_1 h_1^T) h_0 \mu = 1 - \mu h_0^T h_0$$

is also positive definite.

We can continue in a likewise fashion to show that $J_i > 0$ for all $i \geq 0$. We have thus shown that if (3) is satisfied, then $\gamma_{opt} = 1$ and the LMS algorithm achieves it.

■

2.3 Example

To illustrate the robustness of the LMS algorithm we consider a special case of model (1) where h_i is a scalar that randomly takes on the values $+1$ and -1.

$$d_i = h_i w + v_i \tag{8}$$

Suppose we use the LMS algorithm of Theorem 1 with $w_{-1} = 0$ to estimate the weight vector w. Assuming we have observed N points of data, some algebra shows that the transform operator $T_{lms,N}(\mu)$ that maps the disturbances $\{\mu^{-\frac{1}{2}} w, v_i\}$ to the prediction errors $\{e_i\}$ has the following form,

$$\begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_{N-1} \end{bmatrix} =$$

$$\underbrace{\begin{bmatrix} \mu^{\frac{1}{2}} h_0 & 0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}} \alpha h_1 & -\mu h_1 h_0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}} \alpha^2 h_2 & -\mu \alpha h_2 h_0 & -\mu h_2 h_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu^{\frac{1}{2}} \alpha^{N-1} h_{N-1} & -\mu \alpha^{N-2} h_{N-1} h_0 & -\mu \alpha^{N-3} h_{N-1} h_1 & \dots & -\mu h_{N-1} h_{N-2} \end{bmatrix}}_{T_{lms,N}(\mu)} \begin{bmatrix} \mu^{-\frac{1}{2}} w \\ v_0 \\ \vdots \\ v_{N-2} \end{bmatrix} \tag{9}$$

where we have defined $\alpha = 1 - \mu$.

Suppose now we use the RLS algorithm (*viz.* Kalman filter) to estimate the weight vector in (8), *i.e.*,

$$w_{i+1} = w_i + k_{p,i}(d_i - h_i w_i), \qquad w_{-1} = 0$$

where

$$k_{p,i} = \frac{p_i h_i}{1 + p_i |h_i|^2} \quad ,$$

and

$$p_{i+1} = p_i - \frac{|h_i|^2 p_i^2}{1 + p_i |h_i|^2} \quad , \qquad p_0 = \mu \tag{10}$$

At each iteration i this yields the *exact* minimizing solution of

$$\min_w \left[\mu^{-1} |w|^2 + \sum_{j=0}^{i} |d_j - h_j w|^2 \right].$$

Some algebra will also show that we can write the transfer operator $T_{rls,N}$ that maps the disturbances to the prediction errors as follows.

$$\begin{bmatrix} e'_{p,0} \\ e'_{p,1} \\ \vdots \\ e'_{p,N-1} \end{bmatrix} =$$

$$\underbrace{\begin{bmatrix} \mu^{\frac{1}{2}} h_0 & 0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}} \frac{h_1}{1+\mu} & -\mu \frac{h_1 h_0}{1+\mu} & 0 & \dots & 0 \\ \mu^{\frac{1}{2}} \frac{h_2}{1+2\mu} & -\mu \frac{h_2 h_0}{1+2\mu} & -\mu \frac{h_2 h_1}{1+2\mu} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu^{\frac{1}{2}} \frac{h_{N-1}}{1+(N-1)\mu} & -\mu \frac{h_{N-1} h_0}{1+(N-1)\mu} & -\mu \frac{h_{N-1} h_1}{1+(N-1)\mu} & \dots & -\mu \frac{h_{N-1} h_{N-2}}{1+(N-1)\mu} \end{bmatrix}}_{T_{rls,N}(\mu)} \begin{bmatrix} \mu^{-\frac{1}{2}} x_0 \\ v_0 \\ \vdots \\ v_{N-2} \end{bmatrix} \tag{11}$$

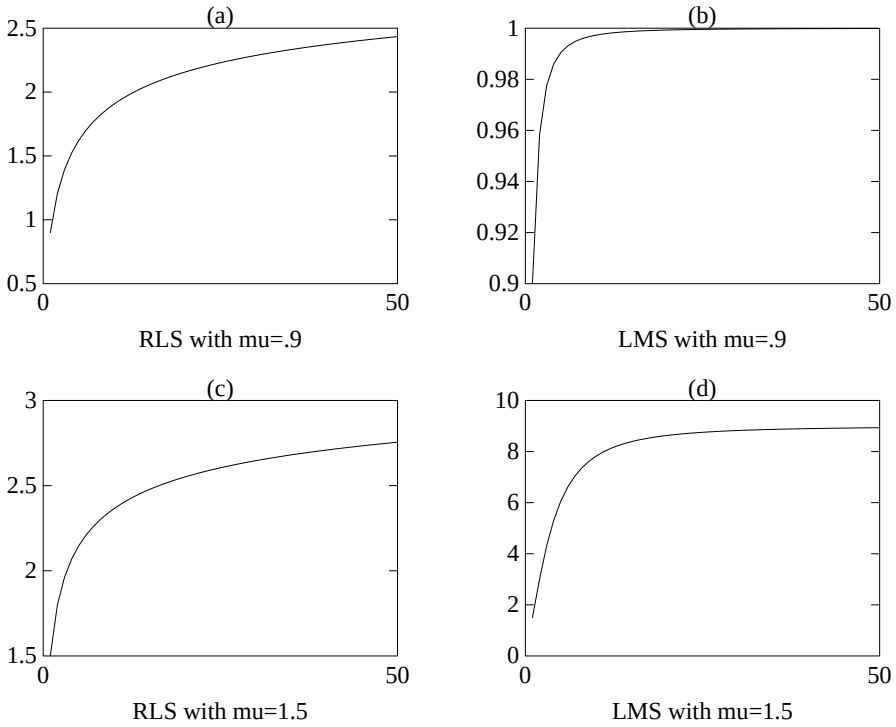

Figure 2: Maximum singular value of transfer operators $T_{lms,N}(\mu)$ and $T_{rls,N}(\mu)$ as a function of N for the values $\mu = .9$ and $\mu = 1.5$.

We now study the maximum singular values of $T_{lms,N}(\mu)$ and $T_{rls,N}(\mu)$ as a function of μ and N. Note that in this special problem, condition (3) implies that μ must be less than one to guarantee the H^∞ optimality of LMS. Therefore we chose the two values $\mu = .9$ and $\mu = 1.5$ (one greater and one less than $\mu = 1$). The results are illustrated in Figure 2 where the maximum singular values of $T_{lms,N}(\mu)$ and $T_{rls,N}(\mu)$ are plotted against the number of observations N. As expected, for $\mu = .9$ the maximum singular value of $T_{lms,N}(\mu)$ remains constant at one, whereas the maximum singular value of $T_{rls,N}(\mu)$ is greater than one and increases with N. For $\mu = 1.5$ both RLS and LMS display maximum singular values greater than one, with the performance of LMS being significantly worse. This justifies the fact that LMS behaves poorly if the learning rate is chosen too large.

Figure 3 shows the worst case disturbance signals for the RLS and LMS algorithms in the $\mu = .9$ case, and the corresponding predicted errors. These worst case disturbances are found by computing the maximum singular vectors of $T_{rls,50}(.9)$ and $T_{lms,50}(.9)$, respectively. The worst case RLS disturbance, and the uncorrupted output $h_i w$, are depicted in Figure 3a. As can be seen from Figure 3b the corresponding RLS predicted error does not go to zero (it is actually biased), whereas the LMS predicted error does. The worst case LMS disturbance signal is given in Figure 3c, and as before, the LMS predicted error tends to zero, while the RLS predicted error does not. The form of the worst case disturbances (especially for RLS) are quite interesting; they compete with the true output early on, and then go to zero.

The disturbance signals considered in this example are rather contrived and may not happen in practice. However, they serve to illustrate the fact that the RLS algorithm may have poor performance even if the disturbance signals have small energy. On the other hand, LMS will have robust performance over a wide range of disturbance signals.

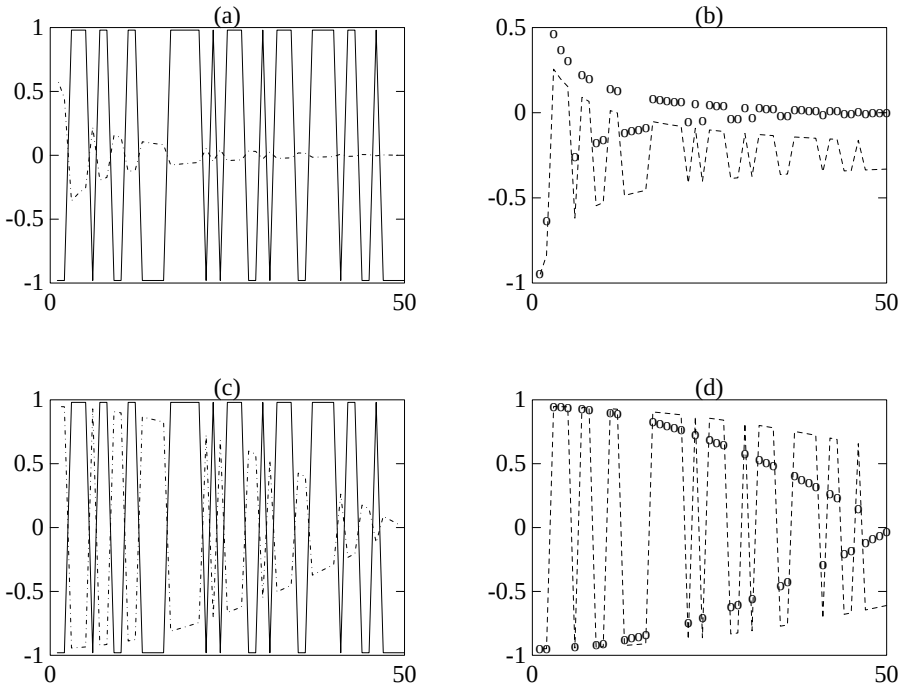

Figure 3: Worst case disturbances and the corresponding predicted errors for RLS and LMS. (a) Solid line represents the uncorrupted output $h_i w$ and the dashed line represents the worst case RLS disturbance. (b) The dashed line and the dotted line represent the RLS and LMS predicted errors, respectively, for the worst case RLS disturbance. (c) Solid line represents the uncorrupted output $h_i w$ and the dashed line represents the worst case LMS disturbance. (d) The dashed line and the dotted line represent the RLS and LMS predicted errors, respectively, for the worst case LMS disturbance.

2.4 Further Remarks

We should mention that the LMS algorithm is only one of a family of H^∞ optimal estimators. For a full parametrization of all such estimators the reader is referred to [8]. However, LMS corresponds to what is called the *central* solution, and has the additional properties of being the maximum entropy solution and the risk-sensitive optimal solution [11, 12, 13].

It is interesting to consider the risk-sensitive optimality of the LMS algorithm in more detail since it will provide LMS with an interpretation in the special case when the disturbances are independent Gaussian random variables. Recall that in the model we have been considering,

$$d_i = h_i w + v_i$$

the disturbances $w - w_{-1}$ and $\{v_i\}$ were assumed to be arbitrary. If we assume them to be zero mean independent Gaussian random variables with covariances μI and 1 respectively, we have the following result.

Theorem 2 (LMS and Risk-sensitivity) *Consider the model (1) where the $w - w_{-1}$ and $\{v_j\}$ are zero mean independent Gaussian random variables with variances μI and 1, respectively. Suppose, moreover, that the $\{h_i\}$ are exciting, and that*

$$0 < \mu < \inf_i \frac{1}{h_i^T h_i}$$

Then the solution to the following minimization problem

$$\min_{\{w_j\}} \left[E exp(\frac{1}{2}\mathbf{C}_i) \right] \tag{12}$$

where $\mathbf{C}_i = \sum_{j=0}^{i} |h_j^T w - h_j^T w_{j-1}|^2$, and the expectation is taken over w and $\{v_j\}$ subject to observing $\{d_0, d_1, \ldots, d_{i-1}\}$, is given by the LMS algorithm

$$w_i = w_{i-1} + \mu h_i (d_i - h_i^T w_{i-1}) \quad , \quad w_{|-1} \tag{13}$$

■

Some intuition concerning this result can be obtained if we consider the cost function appearing in (12),

$$E exp(\frac{1}{2}\mathbf{C}_i).$$

This cost function is a convex and increasing function of $\mathbf{C}_i$. Such a criterion is termed *risk-averse* (or pessimistic) since large weights are on large values of $\mathbf{C}_i$, and hence we are more concerned with the occasional occurence of large values than with the frequent occurence of moderate ones. Thus LMS is a risk-averse filter that avoids the occasional occurence of large prediction error energies, at the expense of admitting the frequent occurence of moderate values of prediction error energy. This is in accordance with the minimax interpretation that we have already established.

Before closing this section we should mention that if instead of the prediction error one were to consider the filtered error $e_{f,i} = h_i w - h_i w_i$, then the H^∞ optimal estimator is the so-called normalized LMS algorithm [8]. The proof follows the same line of reasoning that was pursued in the case of LMS. We state the final result for the sake of completeness.

Theorem 3 (Normalized LMS Algorithm) *Consider the model (1), and suppose we want to minimize the H^∞ norm of the transfer operator from the unknowns $w - w_{-1}$ and $\{v_j\}_{j=0}^i$ to the filtered error $\{e_{f,j} = h_j w - h_j w_j\}_{j=0}^i$. If the input data $\{h_j\}$ is exciting, then the minimum H^∞ norm is $\gamma_{opt} = 1$. In this case an optimal H^∞ estimator is given by the normalized LMS algorithm with parameter μ, viz.*

$$w_{i+1} = w_i + \frac{\mu h_{i+1}}{1 + \mu |h_{i+1}|^2}(d_{i+1} - h_{i+1}^T w_j) \quad , \quad w_{-1} \tag{14}$$

■

3 Nonlinear Minimax Adaptive Filtering

In this section we assume that the observed sequence $\{d_i\}$ obeys the following nonlinear model

$$d_i = h_i(w) + v_i \tag{15}$$

where $h_i(.)$ is a known *nonlinear* function (with bounded first and second order derivatives), w is an unknown weight vector, and $\{v_i\}$ is an unknown disturbance sequence that includes noise and/or modelling errors.

In a neural network context the index i in $h_i(.)$ will correspond to the nonlinear function that maps the weight vector to the output when the ith input pattern is presented, *i.e.*,

$$h_i(w) = h(x^{(i)}, w)$$

where $x^{(i)}$ is the ith input pattern.

As before, we shall denote by $w_i = \mathcal{F}(d_0, \ldots, d_i)$ the estimate of the weight vector using measurements up to and including time i, and the prediction error by

$$e_i = h_i(w) - h_i(w_{i-1})$$

Let T denote the transfer operator that maps the unknowns/disurbances $\{w - w_{-1}, v_i\}$ to the prediction error e_i.

Problem 2 (Optimal Nonlinear Minimax Adaptive Problem) *Find an H^∞-optimal estimation strategy $w_i = \mathcal{F}(d_0, d_1, \ldots, d_i)$ that minimizes $\|T\|_\infty$, and obtain the resulting*

$$\gamma_0^2 = \inf_{\mathcal{F}} \quad \|T\|_\infty^2 = \inf_{\mathcal{F}} \quad \sup_{w,v \in h_2} \quad \frac{\|e\|_2^2}{\mu^{-1}|w - w_{-1}|^2 + \|v\|_2^2} \tag{16}$$

■

Expression (16) clearly states the minimax nature of Problem 2 where, as before, the infimum is taken over all *causal* estimators $\mathcal{F}$.

Currently there is no general solution to the above problem, and the class of nonlinear functions $h_i(.)$ for which there exists a solution is not known [14].

To make some headway, though, note that by using the mean value theorem, expression (15) can be rewritten as

$$d_i = h_i(w_{i-1}) + \frac{\partial h_i}{\partial w}^T (w_{i-1}^*).(w - w_{i-1}) + v_i \tag{17}$$

where w_{i-1}^* is a point on the line connecting w and w_{i-1}, and where $\frac{\partial h_i}{\partial w}^T (w_{i-1}^*)$ denotes the gradient of $h_i(w)$ calculated at this point. We can rearrange (17) as follows

$$d_i - h_i(w_{i-1}) + \frac{\partial h_i}{\partial w}^T (w_{i-1}^*).w_{i-1} = \frac{\partial h_i}{\partial w}^T (w_{i-1}^*).w + v_i \tag{18}$$

If we consider $d_i^{'} = d_i - h_i(w_{i-1}) + \frac{\partial h_i}{\partial w}^T(w_{i-1}^*).w_{i-1}$ to be the new observed sequence, we can apply the LMS algorithm to (18) and obtain the following recursion for w_i

$$\begin{aligned} w_i &= w_{i-1} + \mu \frac{\partial h_i}{\partial w}^T(w_{i-1}^*) \left[d_i^{'} - \frac{\partial h_i}{\partial w}^T(w_{i-1}^*).w_{i-1} \right] \\ &= w_{i-1} + \mu \frac{\partial h_i}{\partial w}^T(w_{i-1}^*)(d_i - h_i(w_{i-1})) \end{aligned} \tag{19}$$

Theorem 1 implies that (19) will yield $\gamma = 1$. The problem with the above algorithm is that the w_i^*'s are not known. But it suggests that the γ_{opt} in Problem 2 (if it exists) cannot be less than one. Moreover, it can be seen that the backpropagation algorithm is an approximation to (19) where w_i^* is replaced by w_i.

To pursue this point further we use again the mean value theorem to write (15) in the alternative form

$$d_i = h_i(w_{i-1}) + \frac{\partial h_i}{\partial w}^T(w_{i-1}).(w - w_{i-1}) + \frac{1}{2}(w - w_{i-1})^T.\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1}) + v_i \tag{20}$$

where once more $\bar{w}_{i-1}$ lies on the line connecting w_{i-1} and w, and where now $\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1})$ denotes the Hessian of $h_i(w)$ calculated at this point. Using (20) and Theorem 1 we have the following result.

Theorem 4 (Backpropagation Algorithm) *Consider the model (15) and the backpropagation algorithm*

$$w_i = w_{i-1} + \mu \frac{\partial h_i}{\partial w}(w_{i-1})(d_i - h_i(w_{i-1})) \qquad , \qquad w_{-1} \tag{21}$$

then if the $\frac{\partial h_i}{\partial w}(w_{i-1})$ *are exciting and*

$$0 < \mu < \inf_i \frac{1}{\frac{\partial h_i}{\partial w}^T(w_{i-1}).\frac{\partial h_i}{\partial w}(w_{i-1})} \tag{22}$$

then for all nonzero w, $v \in h_2$*:*

$$\frac{\| \frac{\partial h_i}{\partial w}^T(w_{i-1}).(w - w_{i-1}) \|_2^2}{\mu^{-1}|w - w_{-1}|^2 + \| v_i + \frac{1}{2}(w - w_{i-1})^T \frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1}) \|_2^2} < 1$$

where

$$(w - w_{i-1})^T.\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1}) = h_i(w) - h_i(w_{i-1}) - \frac{\partial h_i}{\partial w}^T(w_{i-1}).(w - w_{i-1})$$

■

To gain some understanding regarding the above result, consider the new disturbance

$$v_i^{'} = v_i + \frac{1}{2}(w - w_{i-1})^T.\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1})$$

where the second term in $v_i^{'}$ indicates how far $h_i(w)$ is from a *first order approximation* at the point w_{i-1}, and the new prediction error

$$e_i^{'} = \frac{\partial h_i}{\partial w}^T(w_{i-1}).(w - w_{i-1})$$

which is a linearized version of the original prediction error e_i. Theorem 4 implies that the backpropagation algorithm minimizes the H^∞ norm of the transfer operator from the new disturbances $\{w - w_{-1}, v_i^{'}\}$ to the new prediction error $e_i^{'}$. In

particular, backpropagation guarantees that this prediction error energy does not exceed the energy of the new disturbances $w - w_{-1}$ and $v_i^{'}$.

Proofof Theorem 4: Rearrange (20) as follows

$$d_i^{'} = d_i - h_i(w_{i-1}) + \frac{\partial h_i}{\partial w}^T (w_{i-1}).w_{i-1} = \frac{\partial h_i}{\partial w}^T (w_{i-1}).w + v_i^{'} \qquad (23)$$

and apply the LMS algorithm to (23)

$$\begin{aligned} w_i &= w_{i-1} + \mu \frac{\partial h_i}{\partial w}^T (w_{i-1}) \left[d_i^{'} - \frac{\partial h_i}{\partial w}^T (w_{i-1}).w_{i-1} \right] \\ &= w_{i-1} + \mu \frac{\partial h_i}{\partial w}^T (w_{i-1})(d_i - h_i(w_i - 1)) \end{aligned}$$

which is the backpropagation algorithm. Since condition (22) is satisfied, the LMS algorithm applied to (23) guarantees that the energy of the prediction error

$$e_i^{'} = \frac{\partial h_i}{\partial w}^T (w_{i-1}).w - \frac{\partial h_i}{\partial w}^T (w_{i-1}).w_{i-1}$$

never exceeds the energy of the disturbances $\{w - w_{-1}, v_i^{'}\}$.

■

By virtue of Theorem 4 we have been able to show that the backpropagation algorithm minimizes the ratio of the energies of the newly defined prediction errors and disturbances. However, we are really interested in whether backpropagation yields a similar bound on the ratio of the original prediction errors and disturbances. It seems plausible that if w_{-1} is close enough to w then the second term in $v_i^{'}$ should be small, and we should be able to bound the ratio in (16). Thus the following result is expected.

Definition 2 (Persistently Exciting Inputs) *A sequence of input vectors $\{h_i\}$ is called persistently exciting if, and only if,*

$$\lim_{N \to \infty} a^T \left[\sum_{i=0}^{N} h_i h_i^T \right] a = \infty$$

for all $a \in \mathcal{R}^n$. ■

Theorem 5 (Local H^∞ Optimality) *Consider the model (15) and the backpropagation algorithm (21). Suppose that the $\frac{\partial h_i}{\partial w}(w_{i-1})$ are persistently exciting, and that (22) is satisfied. Then for each $\epsilon > 0$, there exist $\delta_1, \delta_2 > 0$ such that for all $|w - w_{-1}| < \delta_1$ and all $v \in h_2$ with $|v_i| < \delta_2$, we have*

$$\frac{\| e \|_2^2}{\mu^{-1}|w - w_{-1}|^2 + \| v \|_2^2} < 1 + \epsilon$$

■

The above Theorem indicates that the backpropagation algorithm is locally H^∞ optimal. In other words for w_{-1} sufficiently close to w, and for sufficiently small disturbance, the ratio in (16) can be made arbitrarily close to one. Note that the conditions on w and v_i are reasonable, since if for example w is too far from w_{-1}, or if some v_i is too large, then it is well known that backpropagation may get stuck in a local minimum, in which case the ratio in (16) may get arbitrarily large.

As before, (22) gives an upper bound on the learning rate μ, and indicates why backpropagation behaves poorly if the learning rate is too large.

Proofof Theorem 5: We shall give an outline of the proof. From Theorem 4 recall that

$$\frac{\| \frac{\partial h_i}{\partial w}(w_{i-1})^T.(w - w_{i-1}) \|_2^2}{\mu^{-1}|w - w_{-1}|^2 + \| v_i + (w - w_{i-1})^T.\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1}) \|_2^2} < 1 \qquad (24)$$

Now we have,

$$\frac{\partial h_i}{\partial w}(w_{i-1})^T.(w - w_{i-1}) = h_i(w) - h_i(w_{-1}) - (w - w_{i-1})^T.\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1})$$

so that if we define $s_i = (w - w_{i-1})^T.\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1})$, we can write

$$\frac{\partial h_i}{\partial w}(w_{i-1})^T.(w - w_{i-1}) = e_i - s_i$$

and (24) becomes

$$\frac{\| e - s \|_2^2}{\mu^{-1}|w - w_{-1}|^2 + \| v + s \|_2^2} < 1$$

Since the denominator is nonzero, we may write

$$\mu^{-1}|w - w_{-1}|^2 + \| v + s \|_2^2 - \| e - s \|_2^2 > 0 \qquad (25)$$

But

$$\begin{aligned}
\| v + s \|_2^2 - \| e - s \|_2^2 &= \sum_{i=0}^{\infty}(v_i + s_i)^2 - \sum_{i=0}^{\infty}(e_i - s_i)^2 \\
&= \sum_{i=0}^{\infty}\left[v_i^2 - e_i^2 + 2(v_i + e_i)s_i\right] \\
&= \| v \|_2^2 - \| e \|_2^2 + \sum_{i=0}^{\infty} 2(v_i + e_i)s_i \\
&\leq \| v \|_2^2 - \| e \|_2^2 + 2 \sup_i |v_i + e_i|. \\
&\quad \sum_{i=0}^{\infty} |(w - w_{i-1})^T.\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1}).(w - w_{i-1})| \\
&\leq \| v \|_2^2 - \| e \|_2^2 + 2\bar{\sigma} \sup_i |v_i + e_i| \sum_{i=0}^{\infty} |w - w_{i-1}|^2
\end{aligned}$$

where $\bar{\sigma} = \sup_i \sigma_{max}\left[\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1})\right]$ is the supremum of the largest singular value of the $\{\frac{\partial^2 h_i}{\partial w^2}(\bar{w}_{i-1})\}$, and is finite since we have assumed bounded second order derivatives. Combining (25) with the last of the above expressions yieds

$$\mu^{-1}|w - w_{-1}|^2 + \| v \|_2^2 - \| e \|_2^2 + 2\bar{\sigma} \sup_i |v_i + e_i| \sum_{i=0}^{\infty} |w - w_{i-1}|^2 > 0$$

from which we conclude

$$\frac{\| e \|_2^2}{\mu^{-1}|w - w_{-1}|^2 + \| v \|_2^2} < 1 + \frac{2\bar{\sigma} \sup_i |v_i + e_i| \sum_{i=0}^{\infty} |w - w_{i-1}|^2}{\mu^{-1}|w - w_{-1}|^2 + \| v \|_2^2} \qquad (26)$$

We now show, under the hypotheses of the Theorem, how to find $\delta_1, \delta_2 > 0$ such that the right hand side of the above inequality is bounded by $1 + \epsilon$. Note that we can construct the following state-space model for the estimation error of the weight vector (*viz.* $w - w_{i-1}$),

$$\begin{aligned} w - w_i &= w - w_{i-1} - \mu \frac{\partial h_i}{\partial w}(w_{i-1})(w_{i-1})(d_i - h_i(w_{i-1})) \\ &= w - w_{i-1} - \mu \frac{\partial h_i}{\partial w}(w_{i-1})(w_{i-1})(h_i(w) - h_i(w_{i-1})) - \\ &\quad \mu \frac{\partial h_i}{\partial w}(w_{i-1}) v_i \end{aligned} \tag{27}$$

If we linearize (27) around $w - w_{i-1} = 0$ and $v_i = 0$, we obtain the following state-space model,

$$\tilde{w}_i = \left[I - \mu \frac{\partial h_i}{\partial w}(w_{i-1}) \frac{\partial h_i}{\partial w}^T (w_{i-1}) \right] \tilde{w}_{i-1} \tag{28}$$

Since the $\{\frac{\partial h_i}{\partial w}(w_{i-1})\}$ are persistently exciting, we can prove that the above linear time-varying system is globally asymptotically stable by verifying that

$$\lim_{N \to \infty} \prod_{i=0}^{N} \left[I - \mu \frac{\partial h_i}{\partial w}(w_{i-1}) \frac{\partial h_i}{\partial w}^T (w_{i-1}) \right] = 0$$

This is a standard result (see *e.g.* [15]). The global asymptotic stability of the linearized model (28) implies the local asymptotic stability of the nonlinear model (27). In other words, there exist $\bar{\delta}_1, \bar{\delta}_2 > 0$ such that for all $|w - w_{-1}| < \bar{\delta}_1$ and all $v \in h_2$ with $|v_i| < \bar{\delta}_2$ the system (27) is asymptotically stable. In particular, for such $\{w - w_{-1}, v_i\}$ the system will have finite energy gain,

$$\frac{\sum_{i=0}^{\infty} |w - w_{i-1}|^2}{\| v \|_2^2} < k_e \tag{29}$$

and finite peak gain,

$$\frac{\sup_i |e_i|}{\sup_i |v_i|} < k_p \tag{30}$$

Using (29) and (30) in (26) implies that for all $|w - w_{-1}| < \bar{\delta}_1$ and all $v \in h_2$ with $|v_i| < \bar{\delta}_2$,

$$\begin{aligned} \frac{\| e \|_2^2}{\mu^{-1}|w - w_{i-1}|^2 + \| v \|_2^2} &< 1 + \frac{2\bar{\sigma}(1 + k_p) \sup_i |v_i| k_e \| v \|_2^2}{\mu^{-1}|w - w_{i-1}|^2 + \| v \|_2^2} \\ &< 1 + 2\bar{\sigma}(1 + k_p) k_e \sup_i |v_i| \end{aligned}$$

If we now choose $\delta_1 = \bar{\delta}_1$ and $\delta_2 = min(\bar{\delta}_2, \frac{\epsilon}{2\bar{\sigma}(1+k_p)k_e})$, we have for all $|w - w_{-1}| < \delta_1$ and all $v \in h_2$ with $|v_i| < \delta_2$,

$$\frac{\| e \|_2^2}{\mu^{-1}|w - w_{i-1}|^2 + \| v \|_2^2} < 1 + \epsilon$$

■

If there is no disturbance in (15) we have the following result.

Corollary 1 *If in addition to the assumptions in Theorem 5 there is no disturbance in (15), then for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $|w - w_{-1}| < \delta$, the backpropagation algorithm will yield $\| e' \|_2^2 \leq \mu^{-1}\delta(1 + \epsilon)$, meaning that the prediction error converges to zero. Moreover w_i will converge to w.* ■

Moreover, provided (22) is satisfied, the larger μ is the faster the convergence will be.

4 Discussion and Conclusion

The results presented in this paper give some new insights into the behaviour of instantaneous gradient-based adaptive algorithms. We showed that if the underlying observation model is linear then LMS is an H^∞ optimal estimator, whereas if the underlying observation model is nonlinear then the backpropagation algorithm is locally H^∞ optimal. The H^∞ optimality of these algorithms explains their inherent robustness to unknown disturbances and modelling errors, as opposed to other estimation algorithms for which such bounds are not guaranteed.

Note that if one considers the transfer operator from the disturbances to the prediction error, then LMS (backpropagation) is H^∞ optimal (locally), over all *causal* estimators. This indicates that our result is most applicable in situations where one is confronted with real-time data and there is no possiblity of storing the training patterns. Such cases arise when one uses adaptive filters or adaptive neural networks for adaptive noise cancellation, channel equalization, real-time control, and undoubtedly many other situations. This is as opposed to pattern recognition, where one has a set of training patterns and repeatedly retrains the network until a desired performance is reached.

In conclusion these results give a new interpretation of the LMS and backpropagation algorithms, which we believe should be worthy of further scrutiny.

References

[1] B. Widrow and M. E. Hoff, Jr. Adaptive switching circuits. *IRE WESCON Conv. Rec.*, pages 96–104, 1960. Pt. 4.

[2] S. Haykin. *Adaptive Filter Theory.* Prentice Hall, Englewood Cliffs, NJ, second edition, 1991.

[3] P. E. Werbos *New methods for ...*

[4] A. Parker *Something new*

[5] D. E. Rumelhart, J. L. McClelland and the PDP Research Group. *Parallel distributed processing : explorations in the microstructure of cognition* Cambridge, Mass. : MIT Press, 1986.

[6] G. Zames. Feedback optimal sensitivity: model preference transformation, multiplicative seminorms and approximate inverses. *IEEE Trans. on Automatic Control*, AC-26:301–320, 1981.

[7] P.P. Khargonekar and K. M. Nagpal. Filtering and smoothing in an $H^\infty-$setting. *IEEE Trans. on Automatic Control*, AC-36:151–166, 1991.

[8] B. Hassibi, A. H. Sayed, and T. Kailath. LMS is H^∞ Optimal. To appear in *IEEE CDC*, San Antonio, Texas, 1993.

[9] B. Widrow and S. D. Stearns. *Adaptive Signal Processing.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1985.

[10] T. Kailath *Linear Systems.* Prentice Hall, Englewood Cliffs NJ, 1980.

[11] P. Whittle. *Risk Sensitive Optimal Control.* John Wiley and Sons, New York, 1990.

[12] K. Glover and D. Mustafa. Derivation of the maximum entropy H^∞ controller and a state space formula for its entropy. *Int. J. Control*, 50:899-916, 1989.

[13] B. Hassibi, A. H. Sayed, and T. Kailath. Recursive linear estimation in Krein spaces - Part II: Applications. To appear in *Proc. IEEE Conference on Decision and Control*, San Antonio, TX, Dec. 1993.

[14] J. A. Ball and J. W. Helton. Nonlinear H^∞ control theory for stable plants. *Math. Control Signals Systems*, 5:233-261, 1992.

[15] M. Vidyasagar. *Nonlinear System Analysis*, 2nd ed., Englewood Cliffs, NJ, Prentice-Hall, 1993.