H^∞ Optimality of the LMS Algorithm[1]

Babak Hassibi[2], Ali H. Sayed[3] and Thomas Kailath

Information Systems Laboratory

Stanford University, Stanford CA 94305

August 28, 1995

Abstract

We show that the celebrated LMS (Least-Mean Squares) adaptive algorithm is H^∞ optimal. The LMS algorithm has been long regarded as an approximate solution to either a stochastic or a deterministic least-squares problem, and it essentially amounts to updating the weight vector estimates along the direction of the instantaneous gradient of a quadratic cost function. In this paper we show that LMS can be regarded as the exact solution to a minimization problem in its own right. Namely, we establish that it is a minimax filter: it minimizes the maximum energy gain from the disturbances to the *predicted* errors, while the closely related so-called normalized LMS algorithm minimizes the maximum energy gain from the disturbances to the *filtered* errors. Moreover, since these algorithms are *central* H^∞ filters, they minimize a certain exponential cost function and are thus also risk-sensitive optimal. We discuss the various implications of these results, and show how they provide theoretical justification for the widely observed excellent robustness properties of the LMS filter.

[1]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. This manuscript is submitted for publication with the understanding that the US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of the Air Force Office of Scientific Research or the U.S. Government.

[2]**Contact author**: Information Systems Laboratory, Stanford University, Stanford CA 94305. Phone (415) 723-1538 Fax (415) 723-8473 E-mail: hassibi@rascals.stanford.edu

[3]The work of A.H. Sayed was also supported by a grant from NSF under award no. MIP-9409319. He is with the Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

1 Introduction

Classical methods in estimation theory (such as maximum-likelihood, maximum entropy and least-squares) require a priori knowledge of the statistical properties of the exogenous signals. In many applications, however, one is faced with model uncertainties and lack of statistical information. Therefore, the introduction of the LMS (Least-Mean-Squares) adaptive filter by Widrow and Hoff 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 [1]. Since this pioneering work, adaptive filtering techniques have been widely used to cope with time variations of system parameters and lack of a priori statistical information [2, 3].

The LMS algorithm was originally conceived as an approximate recursive procedure that solves the following least-squares adaptive problem: given a sequence of $1 \times n$ input row vectors $\{h_i\}$, and a corresponding sequence of desired responses $\{d_i\}$, find an estimate of an $n \times 1$ column vector of weights w, such that the sum of squared errors $\sum_{i=0}^{N} |d_i - h_i w|^2$ is minimized. The LMS solution recursively updates estimates of the weight vector along the direction of the *instantaneous gradient* of the squared error.

Algorithms that *exactly* minimize the sum of squared errors, for every value of N, are also known and are generally referred to as recursive least squares (RLS) algorithms (see, e.g., [3, 4]). Although such exact least-squares algorithms have various desirable optimality properties (such as yielding maximum likelihood estimates) under certain statistical assumptions on the signals (such as temporal whiteness and Gaussian disturbances), they are computationally more complex, and are less robust to disturbance variation than the simple LMS algorithm. For example, it has been observed that the LMS algorithm has better tracking capabilities than the RLS algorithm in the presence of nonstationary inputs [3].

In this paper we show that the superior robustness properties of the LMS algorithm are due to the fact that it is a *minimax* algorithm, or more specifically an H^∞ optimal algorithm. We shall define precisely what this means in Section 3. Here we note only that recently, following some pioneering work in robust control theory (see, *e.g.*, [5]) there has been an increasing interest in minimax estimation (see [6]-[13] and the references therein) with the belief that the resulting so-called H^∞ algorithms will be more robust and less sensitive to model uncertainties and parameter variations. The similarity between the objectives of adaptive filtering and H^∞ estimation suggests that there should be some connection between the two, and indeed our result on the H^∞ optimality of the LMS algorithm provides such a connection.

In addition to giving more insight into the inherent robustness of the LMS algorithm and why it has found such wide applicability in a diverse range of problems, our result provides LMS with a rigorous basis and furnishes a minimization criterion that has long been missing. To be more precise, using some well-known results in H^∞ estimation theory, we show that the LMS algorithm is the so-called central *a priori* H^∞-optimal filter, while the closely related normalized LMS algorithm is the central *a posteriori* H^∞-optimal filter.

The H^∞ optimality property of LMS is a *deterministic* characterization of the algorithm. It is also possible to give a *stochastic* characterization of this algorithm under the assumptions of temporal whiteness and Gaussian disturbances. In this case, we show that LMS minimizes the expected value of a certain exponential cost function, and is therefore risk-sensitive optimal (in the sense of Whittle [16]).

It is ironic that the LMS algorithm is not H^2 optimal, contrary to what its name suggests, but that it rather satisfies a minimax criterion. Moreover, in most H^∞ problems, the optimum solution has not been determined in closed form - what is usually determined is a certain type of suboptimal solution. We show, however, that for the adaptive problem at hand, the optimum solution can be determined.

The remainder of the paper is organized as follows. In Sec. 2 we introduce the problem of adaptive filtering and motivate the question of the robustness of estimators. In order to address the robustness question, we introduce the H^∞ approach in Sec. 3 and formulate the H^∞ estimation problem as one that minimizes the maximum energy gain from the disturbances to the estimation errors. Sec. 4 studies the general problem of state-space H^∞ estimation and, in particular, gives expressions for the H^∞ a posteriori and a priori filters, as well as their full parametrization. The main result is given in Sec. 5 where we formulate the H^∞ adaptive filtering problem as a state-space problem and use the results of Sec. 4 to show that the normalized LMS algorithm is the central a posteriori H^∞ optimal adaptive filter, and that if the learning rate is chosen appropriately, LMS is the central a priori H^∞ optimal adaptive filter. In both cases, the LMS and normalized LMS algorithms guarantee that the energy of the estimation errors never exceeds the energy of the disturbances. Sec. 6 then considers a simple example that demonstrates the robustness of LMS compared to RLS, and also briefly discusses the merits of being H^∞-optimal. In Sec. 7 the full parametrization of all H^∞ optimal adaptive filters is given, and in Sec. 8 we show that LMS and normalized LMS have the additional property of being risk-sensitive optimal. Sec. 9 mentions some further results using the approach and ideas of this paper and Sec. 10 provides the conclusion.

2 Adaptive Filtering

As shown in Fig. 1, suppose we observe an output sequence $\{d_i\}$ that obeys the following model:

$$d_i = h_i w + v_i, \quad i \geq 0 \tag{1}$$

where $h_i = \begin{bmatrix} h_{i1} & h_{i2} & \dots & h_{in} \end{bmatrix}$ is a known $1 \times n$ input vector, $w = \begin{bmatrix} w_1 & w_2 & \dots & w_n \end{bmatrix}^T$ is an unknown $n \times 1$ weight vector that we intend to estimate, and v_i is an unknown disturbance, which may also include modelling errors. We shall not make any assumptions on the noise sequence $\{v_i\}$ (such as stationarity, whiteness, Gaussian distributed, etc.). We denote the estimate of the weight vector using all the information available up to time i by

$$\hat{w}_{|i} = \mathcal{F}(d_0, d_1, \dots, d_i; h_0, h_1, \dots h_i).$$

2.1 Least-Squares Methods

There are a variety of choices for $\hat{w}_{|i}$, but the most widely used estimate is one that satisfies the following least-squares (or H^2) criterion:

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

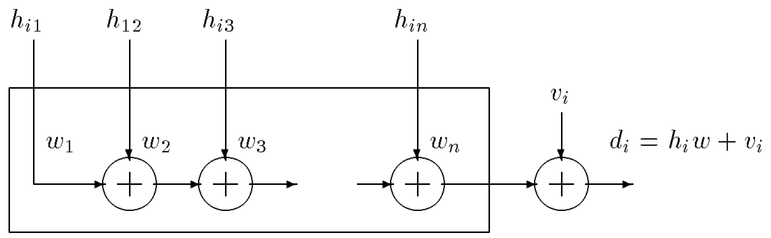

Figure 1: *The model for adaptive filtering.*

where $\hat{w}_{|-1}$ is the initial estimate of w, and $\mu > 0$ represents the relative weight that we give to our initial estimate compared to the "sum of squared-error" term $\sum_{j=0}^{i} |d_j - h_j w|^2$.

The *exact* solution to the above criterion is the RLS (Recursive Least Squares) algorithm:

$$\hat{w}_{|i} = \hat{w}_{|i-1} + k_{p,i}(d_i - h_i \hat{w}_{|i-1}) \quad , \quad \hat{w}_{|-1} \tag{3}$$

with $k_{p,i} = \frac{P_i h_i^*}{1+h_i P_i h_i^*}$ and P_i satisfying the Riccati recursion

$$P_{i+1} = P_i - \frac{P_i h_i^* h_i P_i}{1 + h_i P_i h_i^*}, \quad P_0 = \mu I. \tag{4}$$

The RLS algorithm is used because under suitable stochastic assumptions it has the following two properties:

(a) If $w - \hat{w}_{|-1}$ and the $\{v_j\}$ are assumed to be zero-mean, uncorrelated and, in the case of the $\{v_j\}$, temporally white random variables with variances μI and 1, respectively, then the RLS algorithm minimizes the expected prediction error energy,

$$E \sum_{j=0}^{i} |h_j w - h_j w_{j-1}|^2.$$

(b) If, in addition to the assumptions of part (a), $w - \hat{w}_{|-1}$ and the $\{v_j\}$ are assumed to be jointly Gaussian, then the cost function in (2) becomes the negative of the log-likelihood function and RLS yields the maximum-likelihood estimate of the weight vector w.

2.2 Gradient-Based Methods

In gradient-based algorithms, instead of exactly solving the least-squares problem (2), the estimates of the weight vector are updated along the negative direction of the *instantaneous* gradient of the cost function appearing in (2). Two examples are the LMS (Least-Mean-Squares) [1]

$$\hat{w}_{|i} = \hat{w}_{|i-1} + \mu h_i^*(d_i - h_i \hat{w}_{|i-1}) \quad , \quad \hat{w}_{|-1} \tag{5}$$

and the normalized LMS

$$\hat{w}_{|i} = \hat{w}_{|i-1} + \frac{\mu}{1+\mu h_i h_i^*} h_i^*(d_i - h_i \hat{w}_{|i-1}) \quad , \quad \hat{w}_{|-1} \tag{6}$$

algorithms. Note that in the case of LMS the gain vector $k_{p,i}$ in RLS (which had to be computed by propagating a Riccati equation) has been simply replaced by μh_i^*. Likewise if we compare normalized LMS with the RLS algorithm, we see that the difference is that instead of propagating the matrix P_i via the Riccati recursion we have simply set $P_i = \mu I$, for all i. For this reason the LMS and normalized LMS algorithms have long been considered to be *approximate* least-squares solutions and were thought to lack a rigorous basis.

We should note here that although we have introduced the LMS algorithm as an approximate deterministic least-squares solution, it is also possible to motivate it as an approximate stochastic least-squares solution (see [2, 3]).

2.3 The Question of Robustness

We saw that under suitable stochastic assumptions, the RLS algorithm has certain desirable optimality properties, namely it minimizes the expected prediction error energy and yields maximum-likelihood estimates. However, the question that begs itself is what the performance of such an estimator will be if the assumptions on the disturbances are violated, or if there are modelling errors in our model so that the disturbances must include the modelling errors? In other words

- *is it possible that* **small** *disturbances and modelling errors may lead to* **large** *estimation errors?*

Obviously, a nonrobust algorithm would be one for which the above is true, and a robust algorithm would be one for which small disturbances lead to small estimation errors. More explicitly, in the adaptive filtering problem, where we assume an FIR model, the *true* model may be IIR, but we neglect the tail of the filter response since its components are small. However, unless one uses a robust estimation algorithm, it is conceivable that this small modelling error may result in large estimation errors.

The problem of robust estimation is thus an important one. As we shall see in the next section, the H^∞ estimation formulation is an *attempt* at addressing this question. The idea is to come up with estimators that minimize (or in the suboptimal case, bound) the maximum energy gain from the disturbances to the estimation errors. This will guarantee that if the disturbances are small (in energy) then the estimation errors will be as small as possible (in energy), *no matter what the disturbances are.* In other words the maximum energy gain is minimized over *all possible* disturbances. The robustness of the H^∞ estimators arises from this fact. Since they make no assumption about the disturbances, they have to accomodate for all conceivable disturbances, and are thus over-conservative.

3 The H^∞ Approach

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.

Definition 1 (The H^∞ Norm) *Let h_2 denote the vector space of square-summable complex-valued causal sequences 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 an input sequence $\{u_i\}$ to an output sequence $\{y_i\}$. Then the H^∞ norm of T is defined as*

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

where the notation $\|u\|_2$ denotes the h_2-norm of the causal sequence $\{u_k\}$, viz., $\|u\|_2^2 = \sum_{k=0}^{\infty} u_k^* u_k$.

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

3.1 Formulation of the H^∞ Adaptive Filtering Problem

Recall that $\hat{w}_{|i} = \mathcal{F}(d_0, \ldots d_i; h_0, \ldots h_i)$ denotes the estimate of the weight vector using all the information available from time 0 to time i. In this paper we shall be interested in the following two estimation errors: the *filtered* (or a posteriori) error

$$e_{f,i} = h_i w - h_i \hat{w}_{|i}, \tag{7}$$

and the *predicted* (or a priori) error

$$e_{p,i} = h_i w - h_i \hat{w}_{|i-1}. \tag{8}$$

[Note that in the above errors we compare the estimates $h_i \hat{w}_{|i}$ and $h_i \hat{w}_{|i-1}$ with the *uncorrupted* output $h_i w$ of model (1) and not with the observation d_i.]

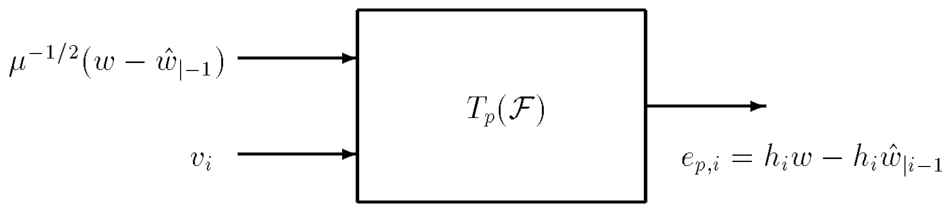

Figure 2: *Transfer operator from the unknown disturbances* $\{\mu^{-1/2}(w - \hat{w}_{|-1}), \{v_j\}_{j=0}^\infty\}$ *to the prediction errors* $\{e_{p,j}\}_{j=0}^\infty$. *Likewise for* $T_f(\mathcal{F})$.

Any choice of estimation strategy $\mathcal{F}(\cdot)$ will induce transfer operators $T_f(\mathcal{F})$ and $T_p(\mathcal{F})$ that map the unknown disturbances $\{\mu^{-1/2}(w - \hat{w}_{|-1}), \{v_j\}_{j=0}^\infty\}$ to the estimation errors $\{e_{f,j}\}_{j=0}^\infty$ and $\{e_{p,j}\}_{j=0}^\infty$, respectively. See Fig. 2.

In the H^∞ framework, robustness is ensured by minimizing the maximum energy gain from the disturbances to the estimation errors. This leads to the following problem.

Problem 1 (H^∞ Adaptive Filtering Problem) *Find an H^∞-optimal estimation strategy $\hat{w}_{|i} = \mathcal{F}_f(d_0, \ldots, d_i; h_0, \ldots, h_i)$, that minimizes $\|T_f(\mathcal{F})\|_\infty$, and an H^∞-optimal strategy $\hat{w}_{|i} = \mathcal{F}_p(d_0, \ldots, d_i; h_0, \ldots, h_i)$, that minimizes $\|T_p(\mathcal{F})\|_\infty$. Also obtain the resulting*

$$\gamma_{f,opt}^2 = \inf_{\mathcal{F}} \; \|T_f(\mathcal{F})\|_\infty^2 = \inf_{\mathcal{F}} \; \sup_{w, v \in h_2} \; \frac{\|e_f\|_2^2}{\mu^{-1}|w - \hat{w}_{|-1}|^2 + \|v\|_2^2}, \tag{9}$$

and

$$\gamma_{p,opt}^2 = \inf_{\mathcal{F}} \ \|T_p(\mathcal{F})\|_\infty^2 = \inf_{\mathcal{F}} \ \sup_{w,v\in h_2} \ \frac{\|e_p\|_2^2}{\mu^{-1}|w-\hat{w}_{|-1}|^2 + \|v\|_2^2}, \tag{10}$$

where $|w-\hat{w}_{|-1}|^2 = (w-\hat{w}_{|-1})^T(w-\hat{w}_{|-1})$.

In order to solve the above H^∞ adaptive filtering problem we shall begin by reviewing some basic results from state-space H^∞ estimation theory. Although it is possible to give a "first principles" derivation of the solution to the above H^∞ adaptive filtering problem (and we shall indeed do so in the Appendix), some study of the more general state-space estimation problem has its own merits, and moreover allows for various generalizations of the results presented here.

4 State-Space H^∞ Estimation

We first give a brief review of some of the results in H^∞ estimation theory using the notation of the companion papers [18, 19]. The reader is also referred to [6]-[13] and the references therein for earlier results and alternative approaches.

4.1 Formulation of the State-Space H^∞ Problem

Consider the time-variant state-space model

$$\begin{cases} x_{i+1} & = & F_i x_i + G_i u_i, & x_0 \\ y_i & = & H_i x_i + v_i, & i \geq 0 \end{cases} \tag{11}$$

where $F_i \in \mathcal{C}^{n\times n}$, $G_i \in \mathcal{C}^{n\times m}$ and $H_i \in \mathcal{C}^{p\times n}$ are known matrices, x_0, $\{u_i\}$, and $\{v_i\}$ are *unknown* quantities and y_i is the measured output. We can regard v_i as a measurement noise and u_i as a process noise or driving disturbance. Let z_i be linearly related to the state x_i via a given matrix $L_i \in \mathcal{C}^{q\times n}$, viz.,

$$z_i = L_i x_i.$$

We shall be interested in the following two cases. Let $\hat{z}_{i|i} = \mathcal{F}_f(y_0, y_1, \ldots, y_i)$ denote an estimate of z_i given observations $\{y_j\}$ from time 0 up to and including time i, and $\hat{z}_i =$

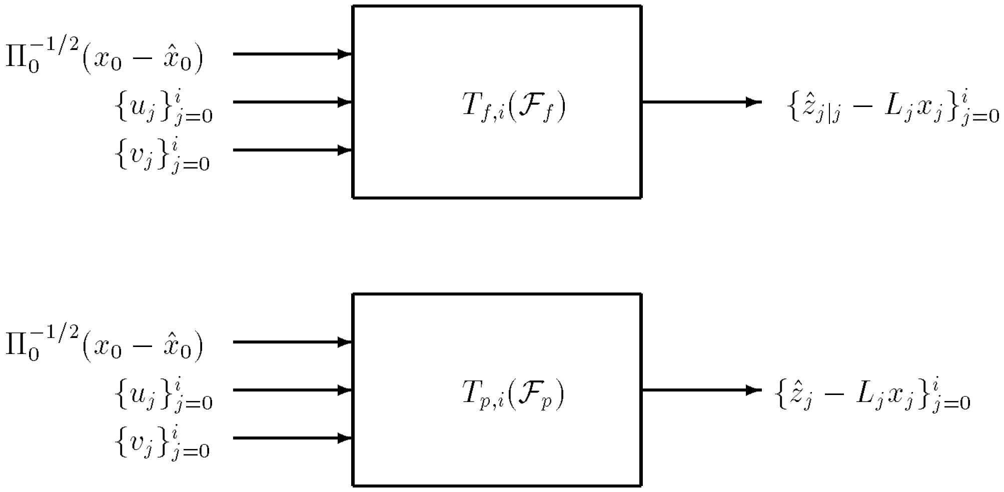

Figure 3: *Transfer matrices from disturbances to filtered and predicted estimation errors.*

$\mathcal{F}_p(y_0, y_1, \ldots, y_{i-1})$ denote an estimate of z_i given observations $\{y_j\}$ from time 0 to time $i-1$. We then have the *filtered* error

$$e_{f,i} = \hat{z}_{i|i} - L_i x_i, \tag{12}$$

and the *predicted* error

$$e_{p,i} = \hat{z}_i - L_i x_i. \tag{13}$$

Let $T_{f,i}(\mathcal{F}_f)$ $(T_{p,i}(\mathcal{F}_p))$ denote the transfer operator that maps the unknown disturbances $\{\Pi_0^{-1/2}(x_0-\hat{x}_0), \{u_j\}_{j=0}^{i}, \{v_j\}_{j=0}^{i}\}$ to the filtered (predicted) errors $\{e_{f,j}\}_{j=0}^{i}$ ($\{e_{p,j}\}_{j=0}^{i}$), where $\hat{x}_0$ denotes an initial guess of x_0, and Π_0 is a given positive definite matrix reflecting a priori knowledge of how close x_0 is to the initial guess $\hat{x}_0$. See Figure 3. The (so-called finite-horizon) H^∞ estimation problem can now be stated as follows.

Problem 2 (Optimal H^∞ Problem) *Find H^∞-optimal estimation strategies $\hat{z}_{i|i} = \mathcal{F}_f(y_0, y_1, \ldots, y_i)$ and $\hat{z}_i = \mathcal{F}_p(y_0, y_1, \ldots, y_{i-1})$ that respectively minimize $\|T_{f,i}(\mathcal{F}_f)\|_\infty$ and $\|T_{p,i}(\mathcal{F}_p)\|_\infty$, and obtain the resulting*

$$\gamma_{f,opt}^2 = \inf_{\mathcal{F}_f} \|T_{f,i}(\mathcal{F}_f)\|_\infty^2 = \inf_{\mathcal{F}_f} \sup_{x_0, u\in h_2, v\in h_2} \frac{\sum_{j=0}^{i} |e_{f,i}|^2}{(x_0-\hat{x}_0)^*\Pi_0{}^{-1}(x_0-\hat{x}_0) + \sum_{j=0}^{i}|u_j|^2 + \sum_{j=0}^{i}|v_j|^2} \tag{14}$$

and

$$\gamma_{p,opt}^2 = \inf_{\mathcal{F}_p} \|T_{p,i}(\mathcal{F}_p)\|_\infty^2 = \inf_{\mathcal{F}_p} \sup_{x_0, u\in h_2, v\in h_2} \frac{\sum_{j=0}^i |e_{p,i}|^2}{(x_0-\hat{x}_0)^*\Pi_0^{-1}(x_0-\hat{x}_0) + \sum_{j=0}^i |u_j|^2 + \sum_{j=0}^i |v_j|^2}. \tag{15}$$

Note that the infimum in (15) is taken over all *strictly* causal estimators $\mathcal{F}_p$, whereas in (14) the estimators $\mathcal{F}_f$ are causal since they have additional access to y_i. This is relevant since the solution to the H^∞ problem, as we shall see, depends on the structure of the information available to the estimator.

The above problem formulation shows that H^∞ optimal estimators guarantee the smallest estimation error energy over all possible disturbances of fixed energy. H^∞ estimators are thus over conservative, which reflects in a better robust behaviour to disturbance variation.

A closed form solution of the optimal H^∞ problem is available only for some special cases (one of which is the adaptive filtering problem as we show here), and a simpler problem results if one relaxes the minimization condition and settles for a suboptimal solution.

Problem 3 (Sub-optimal H^∞ Problem) *Given scalars $\gamma_f > 0$ and $\gamma_p > 0$, find estimation strategies $\hat{z}_{i|i} = \mathcal{F}_f(y_0, y_1, \ldots, y_i)$ and $\hat{z}_i = \mathcal{F}_p(y_0, y_1, \ldots, y_{i-1})$ that respectively achieve $\| T_{f,i}(\mathcal{F}_f) \|_\infty < \gamma_f$ and $\| T_{p,i}(\mathcal{F}_p) \|_\infty < \gamma_p$. This clearly requires checking whether $\gamma_f \geq \gamma_{f,o}$ and $\gamma_p \geq \gamma_{p,o}$.*

The above two problem formulations are for the finite horizon case. In the infinite horizon case, to guarantee that $\|T_f(\mathcal{F})\|_\infty \leq \gamma_f$ and $\|T_p(\mathcal{F})\|_\infty \leq \gamma_p$ we need to ensure $\|T_{f,i}(\mathcal{F})\|_\infty < \gamma_f$ and $\|T_{p,i}(\mathcal{F})\|_\infty < \gamma_p$ for all i.

4.2 The H^∞ Filters

We now briefly review the solutions of the H^∞ filtering problems using the notation of [18, 19].

Theorem 1 (The H^∞ Aposteriori Filter) *For a given $\gamma > 0$, if the F_i are nonsingular then an estimator with $\|T_{f,i}\|_\infty < \gamma$ exists if, and only if,*

$$P_j^{-1} + H_j^* H_j - \gamma^{-2} L_j^* L_j > 0, \qquad j = 0, \ldots, i \tag{16}$$

where $P_0 = \Pi_0$, *and* P_j *satisfies the Riccati recursion*

$$P_{j+1} = F_j P_j F_j^* + G_j G_j^* - \begin{bmatrix} L_j^* & H_j^* \end{bmatrix} R_{e,j}^{-1} \begin{bmatrix} L_j \\ H_j \end{bmatrix} P_j F_j^* \tag{17}$$

with

$$R_{e,j} = \begin{bmatrix} -\gamma^2 I & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} + \begin{bmatrix} L_j \\ H_j \end{bmatrix} P_j \begin{bmatrix} L_j^* & H_j^* \end{bmatrix}.$$

If this is the case, then one possible H_∞ *filter with level* γ *is given by*

$$\hat{z}_{j|j} = L_j \hat{x}_{j|j},$$

where $\hat{x}_{j|j}$ *is recursively computed as*

$$\hat{x}_{j+1|j+1} = F_j \hat{x}_{j|j} + K_{f,j+1}(y_{j+1} - H_{j+1} F_j \hat{x}_{j|j}), \quad \hat{x}_{-1|-1} = \textit{initial guess} \tag{18}$$

and

$$K_{f,j+1} = P_{j+1} H_{j+1}^* (I + H_{j+1} P_{j+1} H_{j+1}^*)^{-1}. \tag{19}$$

Theorem 2 (The H^∞ Apriori Filter) *For a given* $\gamma > 0$, *if the* F_i *are nonsingular then an estimator with* $\|T_{p,i}\|_\infty < \gamma$ *exists if, and only if,*

$$\tilde{P}_j^{-1} = P_j^{-1} - \gamma^{-2} L_j^* L_j > 0, \qquad j = 0, \ldots, i \tag{20}$$

where P_j *is the same as in Theorem 1. If this is the case, then one possible* H_∞ *filter with level* γ *is given by*

$$\hat{z}_j = L_j \hat{x}_j, \tag{21}$$

$$\hat{x}_{j+1} = F_j \hat{x}_j + K_{p,j}(y_j - H_j \hat{x}_j), \quad \hat{x}_0 = \textit{initial guess} \tag{22}$$

where

$$K_{p,j} = F_j \tilde{P}_j H_j^* (I + H_j \tilde{P}_j H_j^*)^{-1}. \tag{23}$$

Note that the above two estimators bear a striking resemblance to the celebrated Kalman filter:

$$\left\{ \begin{array}{lcl} \hat{x}_{j+1} & = & F_j \hat{x}_j + F_j P_j H_j^* (I + H_j P_j H_j^*)^{-1} (y_j - H_j \hat{x}_j) \\ P_{j+1} & = & F_j P_j F_j^* + G_j G_j^* - F_j P_j (I + H_j P_j H_j^*)^{-1} P_j F_j^* \end{array} \right. \tag{24}$$

and that the only difference is that the P_j of equation (19), and $\tilde{P}_j$ of equation (23), satisfy Riccati recursions that differ with that of (24). However, as $\gamma \to \infty$, the Riccati recursion (17) collapses to the Kalman filter recursion (24). This suggests that the H^∞ norm of the Kalman filter may be quite large, indicating that it may have poor robustness properties.

It is also interesting that the structure of the H^∞ estimators depends, via the Riccati recursion (17), on the linear combination of the states that we intend to estimate (*i.e.*, the L_i). This is as opposed to the Kalman filter, where the estimate of any linear combination of the state is given by that linear combination of the state estimate. Intuitively, this means that the H^∞ filters are specifically tuned towards the linear combination $L_i x_i$.

Note also that condition (20) is more stringent than condition (16), indicating that the existence of an a priori filter of level γ implies the existence of an a posteriori filter of level γ, but not necessarily vice versa.

We further remark that the filter of Theorem 1 (and Theorem 2) is one of many possible filters with level γ. A full parametrization of all estimators of level γ are given by the following Theorems. (For proofs see [19]).

Theorem 3 (All H^∞ Aposteriori Estimators) *All H^∞ a posteriori estimators that achieve a level γ_f (assuming they exist) are given by*

$$\begin{aligned} \hat{z}_{j|j} \;=\; & L_j \hat{x}_{j|j} + [\gamma_f^2 I - L_j (P_j^{-1} + H_j^* H_j)^{-1} L_j^*]^{\frac{1}{2}} \\ & \mathcal{S}_j \left((I + H_j P_j H_j^*)^{\frac{1}{2}} (y_j - H_j \hat{x}_{j|j}), \ldots, (I + H_0 P_0 H_0^*)^{\frac{1}{2}} (y_0 - H_0 \hat{x}_{0|0}) \right) \end{aligned} \tag{25}$$

where $\hat{x}_{j|j}$ satisfies the recursion

$$\hat{x}_{j+1|j+1} = F_j \hat{x}_{j|j} + K_{f,j+1} (y_{j+1} - H_{j+1} F_j \hat{x}_{j|j}) - K_{c,j} (\hat{z}_{j|j} - L_j \hat{x}_{j|j}) \tag{26}$$

with $K_{f,j+1}$ the same as in Theorem 1,

$$K_{c,j} = (I + P_{j+1} H_{j+1} H_{j+1}^*)^{-1} F_j (P_j^{-1} + H_j H_j^* - \gamma_f^{-2} L_j L_j^*)^{-1} L_j^*, \tag{27}$$

and

$$\mathcal{S}(a_j,\ldots,a_0) = \begin{bmatrix} \mathcal{S}_0(a_0) \\ \mathcal{S}_1(a_1,a_0) \\ \vdots \\ \mathcal{S}_j(a_j,\ldots,a_0) \end{bmatrix}$$

is any (possibly nonlinear) contractive causal mapping, i.e.,

$$\sum_{j=0}^{k} |\mathcal{S}_j(a_j,\ldots,a_0)|^2 < \sum_{j=0}^{k} |a_j|^2 \quad \textit{for all} \quad k = 0,1,\ldots,i.$$

Theorem 4 (All H^∞ Apriori Estimators) *All H^∞ a priori estimators that achieve a level γ_p (assuming they exist) are given by*

$$\begin{aligned} \hat{z}_j \; &= \; L_j\hat{x}_j + (\gamma_p^2 I - L_j P_j L_j^*)^{\frac{1}{2}} \\ & \mathcal{S}_j\left((I + H_{j-1}\tilde{P}_{j-1}H_{j-1}^*)^{-\frac{1}{2}}(y_{j-1} - H_{j-1}\bar{x}_{j-1}),\ldots,(I + H_0\tilde{P}_0H_0^*)^{-\frac{1}{2}}(y_0 - H_0\bar{x}_0)\right) \end{aligned} \tag{28}$$

where

$$\bar{x}_k = \hat{x}_k + P_k L_k^*(-\gamma_p^2 I + L_k P_k L_k^*)^{-1}(\hat{z}_k - L_k\hat{x}_k), \tag{29}$$

$\hat{x}_j$ satisfies the recursion

$$\hat{x}_{j+1|j} = F_j\hat{x}_{j|j-1} + F_j P_j \begin{bmatrix} L_j^* & H_j^* \end{bmatrix} R_{e,j}^{-1} \begin{bmatrix} \hat{z}_j - L_j\hat{x}_{j|j-1} \\ y_j - H_j\hat{x}_{j|j-1} \end{bmatrix}, \tag{30}$$

with P_j, $\tilde{P}_j$ and $R_{e,j}$ given by Theorem 2, and $\mathcal{S}$ is any (possibly nonlinear) contractive causal mapping.

Note that although the filters obtained in Theorems 1 and 2 are linear, the full parametrization of all H^∞ filters with level γ is given by a *nonlinear* causal contractive mapping $\mathcal{S}$. The filters of Theorems 1 and 2 are known as the *central* filters and correspond to $\mathcal{S} = 0$. These central filters have a number of other interesting properties. They correspond, as we shall see in a subsequent section, to the risk-sensitive optimal filter [16], and can be shown to be the *maximum entropy* filter [21].

5 Main Result

Let us first note that the basic equation of the adaptive filtering model (1) can be rewritten in the following state-space form:

$$\begin{cases} x_{i+1} = x_i \\ d_i = h_i x_i + v_i \end{cases} \qquad x_0 = w. \tag{31}$$

This is a relevant step since it reduces the adaptive filtering problem to an equivalent state-space estimation problem. This point of view has been recently proposed in [4] where a unified square-root-based derivation of exponentially-weighted RLS adaptive algorithms is obtained by reformulating the original adaptive problem as a state-space linear least-squares estimation problem and then applying various algorithms from Kalman filter theory. Here we shall instead apply the H^∞ theory to the state-space model (31) and show that the optimum a priori and a posteriori H^∞ filters reduce to the LMS and normalized LMS algorithms, respectively.

At this point we need one more definition.

Definition 2 (Exciting Inputs) *The input vectors h_i are called exciting if, and only if,*

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

5.1 The Normalized LMS Algorithm

We first consider the a posteriori filter and show that it collapses to the normalized LMS algorithm.

Theorem 5 (Normalized LMS Algorithm) *Consider the state-space model (31), and suppose we want to minimize the H^∞ norm of the transfer operator $T_f(\mathcal{F})$ from the unknowns $\mu^{-1/2}(w - \hat{w}_{|-1})$ and $\{v_j\}_{j=0}^{\infty}$ to the filtered error $\{e_{f,j} = \hat{z}_{j|j} - h_j w\}_{j=0}^{\infty}$. If the input data $\{h_j\}$ is exciting, then the minimum H^∞ norm is*

$$\gamma_{f,opt} = 1.$$

In this case, the central optimal H^∞ a posteriori filter is

$$\hat{z}_{j|j} = h_j \hat{w}_{|j},$$

where $\hat{w}_{|j}$ is given by the normalized LMS algorithm with parameter μ,

$$\hat{w}_{|j+1} = \hat{w}_{|j} + \frac{\mu h_{j+1}^*}{1 + \mu h_{j+1} h_{j+1}^*}(d_{j+1} - h_{j+1}\hat{w}_{|j}), \quad \hat{w}_{|-1} = \textit{initial guess}. \tag{32}$$

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

$$h_i \hat{w}_{|-1} = h_i w + v_i = d_i.$$

In this case one expects that the estimator will not change its estimate of w, so that $\hat{w}_{|i} = \hat{w}_{|-1}$ for all i. Thus the filtered error is

$$e_{f,i} = h_i w - h_i \hat{w}_{|i} = h_i w - h_i \hat{w}_{|-1} = -v_i,$$

and the ratio in (9) becomes

$$\frac{\|v\|^2}{\mu^{-1}|w - \hat{w}_{|-1}|^2 + \|v\|^2} = \frac{\|h_i(w - \hat{w}_{|-1})\|^2}{\mu^{-1}|w - \hat{w}_{|-1}|^2 + \|h_i(w - \hat{w}_{|-1})\|^2}.$$

When the $\{h_i\}$ are exciting, for any $\epsilon > 0$, we can find a weight vector w and an integer N such that $\sum_{i=0}^N |h_i(w - \hat{w}_{|-1})|^2 \geq \frac{|w - \hat{w}_{|-1}|^2}{\epsilon\mu}$. With these choices we have

$$\frac{\sum_{i=0}^N |h_i(w - \hat{w}_{|-1})|^2}{\mu^{-1}|w - \hat{w}_{|-1}|^2 + \sum_{i=0}^N |h_i(w - \hat{w}_{|-1})|^2} \geq 1 - \epsilon,$$

so that the ratio in (9) can be made arbitrarily close to one.

The surprising fact though is that $\gamma_{f,opt}$ is exactly one and that the normalized LMS algorithm achieves it. What this means is that normalized LMS guarantees that the energy of the filtered 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 small energy that will yield a filtered error of large energy [20].

Proof of Theorem 5: We apply the aposteriori filter of Theorem 1 to the state-space model (31) where $F_i = I$, $G_i = 0$, $H_i = h_i$, and $L_i = h_i$. Thus the Riccati equation simplifies to

$$P_{i+1} = P_i - P_i \begin{bmatrix} h_i^* & h_i^* \end{bmatrix} \left\{ \begin{bmatrix} -\gamma^2 I & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} + \begin{bmatrix} h_j \\ h_j \end{bmatrix} P_i \begin{bmatrix} h_i^* & h_i^* \end{bmatrix} \right\}^{-1} \begin{bmatrix} h_i \\ h_i \end{bmatrix} P_i,$$

which, using the matrix inversion lemma [23], implies that

$$\begin{aligned} P_{i+1}^{-1} &= P_i^{-1} + \begin{bmatrix} h_i^* & h_i^* \end{bmatrix} \begin{bmatrix} -\gamma^{-2} I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} h_i \\ h_i \end{bmatrix} \\ &= P_i^{-1} + (1-\gamma^{-2}) h_i^* h_i. \end{aligned}$$

Consequently, starting with $P_0^{-1} = \mu^{-1} I$, we get

$$P_{i+1}^{-1} = \mu^{-1} I + (1-\gamma^{-2}) \sum_{j=0}^{i} h_j^* h_j. \tag{33}$$

Now we need to check the existence condition (16) and find the optimum $\gamma_{f,opt}$. It follows from the above expression for P_{i+1}^{-1} that we have

$$P_{i+1}^{-1} + H_{i+1}^* H_{i+1} - \gamma^{-2} L_{i+1}^* L_{i+1} = \mu^{-1} I + (1-\gamma^{-2}) \sum_{j=0}^{i+1} h_j^* h_j. \tag{34}$$

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 on the right hand side of (34) is negative, viz.,

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

Consequently, $P_{i+1}^{-1} + H_{i+1}^* H_{i+1} - \gamma^{-2} L_{i+1}^* L_{i+1}$ cannot be positive-definite. Therefore, $\gamma_{f,opt} \geq 1$. We now verify that $\gamma_{f,opt}$ is indeed 1. For this purpose, we note that if we consider $\gamma = 1$ then from equation (33) we have $P_i = \mu I > 0$ for all i and the existence condition is satisfied. If we now write the a posteriori filter for $\gamma_{f,opt} = 1$, with $P_i = \mu I$, we get the desired so-called normalized LMS algorithm (32).

∎

5.2 The LMS Algorithm

We now apply the a priori H^∞-filter and show that it collapses to the LMS algorithm.

Theorem 6 (LMS Algorithm) *Consider the state-space model (31), and suppose we want to minimize the H^∞ norm of the transfer operator $T_p(\mathcal{F})$ from the unknowns $\mu^{-1/2}(w - \hat{w}_{|-1})$ and $\{v_j\}_{j=0}^\infty$ to the predicted error $\{e_{p,j} = \check{z}_j - h_j w\}_{j=0}^\infty$. If the input data $\{h_j\}$ is exciting, and*

$$0 < \mu < \inf_i \frac{1}{h_i h_i^*} \tag{35}$$

then the minimum H^∞ norm is

$$\gamma_{p,opt} = 1.$$

In this case, the central optimal a priori H^∞ filter is

$$\check{z}_j = h_i \hat{w}_{|j-1}$$

where $\hat{w}_{|j-1}$ is given by the LMS algorithm with learning rate μ, viz.,

$$\hat{w}_{|j} = \hat{w}_{|j-1} + \mu h_j^*(d_j - h_j \hat{w}_{|j-1}) \quad , \quad \hat{w}_{|-1}. \tag{36}$$

Proof: The proof is similar to that for the normalized LMS case. For $\gamma < 1$, the matrix $\tilde{P}_i$ of Theorem 2 cannot be positive-definite. For $\gamma = 1$, we get $P_i = \mu I > 0$ for all i, and

$$\begin{aligned} \tilde{P}_i^{-1} &= P_i^{-1} - L_i^* L_i \\ &= \mu^{-1} I - h_i^* h_i \end{aligned}$$

It is straightforward to see that the eigenvalues of $\tilde{P}_i^{-1}$ are

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

Thus $\tilde{P}_i^{-1}$ is positive definite if, and only if, (35) is satisfied, which leads to $\gamma_{p,opt} = 1$. Writing the H^∞ a priori filter equations for $\gamma = 1$ yields

$$\hat{w}_{|i} = \hat{w}_{|i-1} + \tilde{P}_i h_i^*(I + h_i \tilde{P}_i h_i^*)^{-1}(d_i - h_i \hat{w}_{|i-1})$$

$$
\begin{aligned}
&= \hat{w}_{|i-1} + \tilde{P}_i(I + h_i^* h_i \tilde{P}_i)^{-1} h_i^*(d_i - h_i \hat{w}_{|i-1}) \\
&= \hat{w}_{|i-1} + (\tilde{P}_i^{-1} + h_i^* h_i)^{-1} h_i^*(d_i - h_i \hat{w}_{|i-1}) \\
&= \hat{w}_{|i-1} + \mu h_i^*(d_i - h_i \hat{w}_{|i-1}).
\end{aligned}
$$

■

The above result indicates that if the learning rate μ is chosen according to (35), then LMS ensures that the energy of the predicted error will never exceed the energy of the disturbances. It is interesting that we have obtained an upper bound on the learning rate μ that guarantees this H^∞ optimality, since it is a well known fact that LMS behaves poorly if the learning rate is chosen too large. It is also interesting to compare the bound in (35) with the bounds studied in [2] and [24].

We further note that if the input data is not exciting, then $\sum_{i=0}^{\infty} h_i^* h_i$ will have a finite limit, and the minimum H^∞ norm of the a posteriori and a priori filters will be the smallest γ that ensures

$$
\mu^{-1} I + (1 - \gamma^{-2}) \sum_{i=0}^{\infty} h_i^* h_i > 0.
$$

This will in general yield $\gamma_{opt} < 1$, and Theorems 1 and 2 can be used to write the optimal filters for this γ_{opt}. In this case the LMS and normalized LMS algorithms will still correspond to $\gamma = 1$, but will now be suboptimal.

6 An Illustrative Example

To illustrate the robustness of the LMS algorithm we consider a special case of model (31), where h_i is now a scalar that randomly takes on the values $+1$ and -1.

Using the LMS algorithm we can write the following state-space model for the predicted error $e_{p,i} = h_i x_i - h_i \hat{x}_i$:

$$
\begin{cases}
\tilde{x}_{i+1} = (1 - \mu |h_i|^2)\tilde{x}_i - \mu h_i^* v_i = (1-\mu)\tilde{x}_i - \mu h_i v_i \\
e_{p,i} = h_i \tilde{x}_i
\end{cases}
, \quad \tilde{x}_0 = w - \hat{x}_{-1} \tag{37}
$$

where $\tilde{x}_i = x_i - \hat{x}_i$, and where we have used the fact that the h_i have magnitude one. Assuming we have observed N points of data, we can then use (37) to write the operator, $T_{lms,N}(\mu)$, that maps the disturbances $\{\mu^{-\frac{1}{2}}\tilde{x}_0, \{v_i\}_{i=0}^{N-1}\}$ to the $\{e_{p,i}\}_{i=0}^{N-1}$.

$$\begin{bmatrix} e_{p,0} \\ e_{p,1} \\ \vdots \\ e_{p,N-1} \end{bmatrix} =$$

$$\underbrace{\begin{bmatrix} \mu^{\frac{1}{2}}h_0 & 0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}}(1-\mu)h_1 & -\mu h_1 h_0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}}(1-\mu)^2 h_2 & -\mu(1-\mu)h_2 h_0 & -\mu h_2 h_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu^{\frac{1}{2}}(1-\mu)^{N-1}h_{N-1} & -\mu(1-\mu)^{N-2}h_{N-1}h_0 & -\mu(1-\mu)^{N-3}h_{N-1}h_1 & \dots & -\mu h_{N-1}h_{N-2} \end{bmatrix}}_{T_{lms,N}(\mu)} \begin{bmatrix} \mu^{-\frac{1}{2}}\tilde{x}_0 \\ v_0 \\ \vdots \\ v_{N-2} \end{bmatrix}. \tag{38}$$

Suppose now we use the RLS algorithm (*viz.* the Kalman filter) to estimate the states in (31), *i.e.*,

$$\hat{x}_{i+1} = \hat{x}_i + k_{p,i}(d_i - h_i\hat{x}_i)$$

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

$$p_{i+1} = p_i - \frac{|h_i|^2 p_i^2}{1+p_i|h_i|^2} = p_i - \frac{p_i^2}{1+p_i} = \frac{p_i}{1+p_i}, \quad p_0 = \mu. \tag{39}$$

Then we may write the following state-space model for the RLS predicted error $e'_{p,i} = h_i x_i - h_i\hat{x}_i$,

$$\left\{ \begin{array}{rcl} \tilde{x}_{i+1} & = & (1-k_{p,i}h_i)\tilde{x}_i - k_{p,i}v_i \\ e'_{p,i} & = & h_i\tilde{x}_i \end{array} \right. , \quad \tilde{x}_0 = w - \hat{x}_{-1} \tag{40}$$

Now solving (39) yields

$$p_i = \frac{\mu}{1+i\mu} \quad , \tag{41}$$

and

$$k_{p,i} = h_i p_{i+1}, \qquad 1 - k_{p,i}h_i = \frac{p_{i+1}}{p_i}. \tag{42}$$

Using (41), (42), and the state-space model (40) we can also write the transfer operator

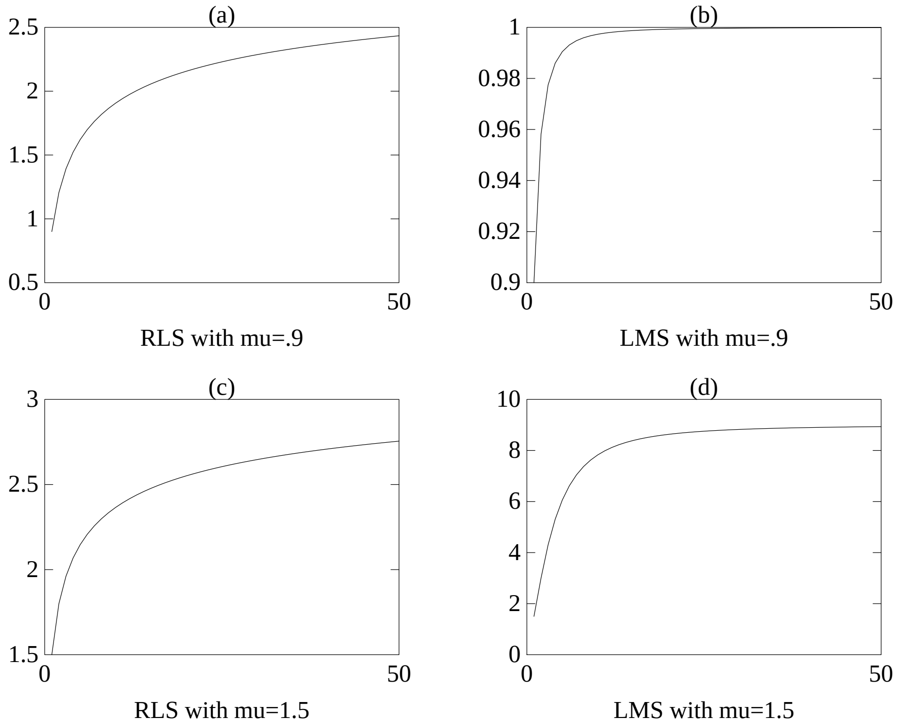

Figure 4: *Maximum singular value of transfer operators $T_{lms,N}(\mu)$ and $T_{rls,N}(\mu)$ as a function of N for the values $\mu = .9$ and $\mu = 1.5$.*

$T_{rls,N}(\mu)$ that maps the disturbances to the predicted errors as follows:

$$\begin{bmatrix} e'_{p,0} \\ e'_{p,1} \\ \vdots \\ e'_{p,N-1} \end{bmatrix} = \underbrace{\begin{bmatrix} \mu^{\frac{1}{2}} h_0 & 0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}} \frac{h_1}{1+\mu} & -\mu \frac{h_1 h_0}{1+\mu} & 0 & \dots & 0 \\ \mu^{\frac{1}{2}} \frac{h_2}{1+2\mu} & -\mu \frac{h_2 h_0}{1+2\mu} & -\mu \frac{h_2 h_1}{1+2\mu} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu^{\frac{1}{2}} \frac{h_{N-1}}{1+(N-1)\mu} & -\mu \frac{h_{N-1} h_0}{1+(N-1)\mu} & -\mu \frac{h_{N-1} h_1}{1+(N-1)\mu} & \dots & -\mu \frac{h_{N-1} h_{N-2}}{1+(N-1)\mu} \end{bmatrix}}_{T_{rls,N}(\mu)} \begin{bmatrix} \mu^{-\frac{1}{2}} x_0 \\ v_0 \\ \vdots \\ v_{N-2} \end{bmatrix}. \tag{43}$$

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 (35) 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 4 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.

Figure 5 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 $h_i x_i$, are depicted in Figure 5a. As can be seen from Figure 5b, 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 5c, 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.

6.1 Discussion

In Section 5.1 we motivated the $\gamma_{f,opt} = 1$ result for normalized LMS by considering a disturbance strategy that made the observed output d_i coincide with the expected output $h_i \hat{w}_{|-1}$. It is now illuminating to consider the *dual* strategy for the estimator.

Recall that in the a posteriori adaptive filtering problem the estimator has access to observations $d_0, d_1, \ldots, d_i$ and is required to construct an estimate of $\hat{z}_{i|i}$ of the uncorrupted output $z_i = h_i x_i$. The dual to the above mentioned disturbance strategy would be to construct an estimate that coincides with the observed output, *viz.*,

$$\hat{z}_{i|i} = d_i. \tag{44}$$

The corresponding filtered error is:

$$e_{f,i} = \hat{z}_{i|i} - h_i x_i = d_i - h_i x_i = v_i.$$

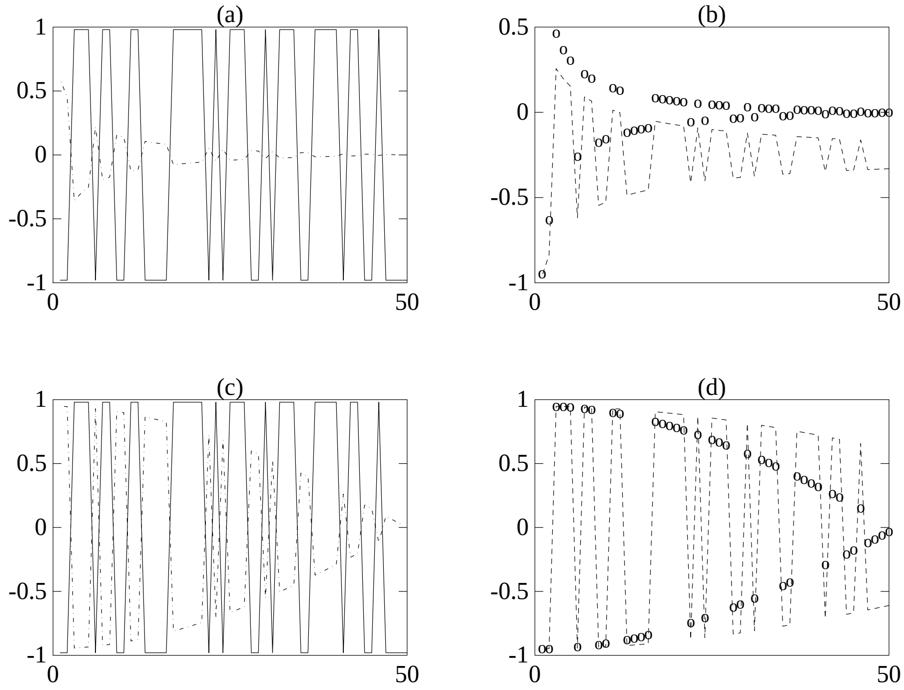

Figure 5: *Worst case disturbances and the corresponding predicted errors for RLS and LMS. (a) The solid line represents the uncorrupted output* $h_i x_i$ *and the dashed line represents the worst case RLS disturbance. (b) The dashed line and the dotted line represent the RLS and LMS predicted errors, respectively, for the worst case RLS disturbance. (c) The solid line represents the uncorrupted output* $h_i x_i$ *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.*

Thus the ratio in (9) can be made arbitrarily close to one, and the estimator (44) will achieve the same $\gamma_{f,opt} = 1$ that the normalized LMS algorithm does.

The fact that the simplistic estimator (44) (which is obviously of no practical use) is an optimal H^∞ aposteriori filter seems to question the very merit of being H^∞ optimal. A first indication towards this direction may be the fact that the H^∞ estimators that achieve a certain level γ are nonunique. In our opinion, the property of being H^∞ optimal (*i.e.* , of minimizing the energy gain from the disturbances to the errors) is a desirable property in itself. The high sensitivity of the RLS algorithm to different disturbance signals, as illustrated in the example of Section 6, clearly indicates the desirability of the H^∞ optimality property. However, different estimators in the set of all H^∞ optimal estimators may have drastically different behaviour with respect to other *desirable* performance measures.

In Section 7 we develop the full parametrization of all H^∞ optimal a posteriori and a priori adaptive filters, and show how to obtain (44) as a special case of this parametrization. Moreover, it can be shown (see [22]) that among all H^∞-optimal a posteriori filters the filter (44) has the worst H^2 (or, roughly speaking, average) performance. Thus it is the least desirable H^∞-optimal filter with respect to an H^2 criterion. On the other hand, as indicated in Theorems 5 and 6, the LMS and normalized LMS algorithms correspond to the so-called central filters. These central filters have other desirable properties that we discuss in Section 8: they are risk-sensitive optimal and can also be shown to be maximum entropy.

The main problem with the estimator (44) is that it makes absolutely no use of the state-space model (31). We should note that it is not possible to come up with such a simple minded estimator in the a priori case: indeed as we shall see in the next section, the a priori estimator corresponding to (44) is highly nontrivial. The reason seems to be that since in the a priori case one deals with predicted error energy, it is inevitable that one must make use of the state-space model (31) in order to construct an optimal prediction of the *next* output. Thus in the a priori case, the problems arising from such unreasonable estimators such as (44) are avoided.

7 All H^∞ Adaptive Filters

In Section 6.1 we came up with an alternative optimal H^∞ a posteriori filter. We now use the results of Theorems 3 and 4 to parametrize all optimal H^∞ a priori and a posteriori filters.

Theorem 7 (All H^∞ Aposteriori Adaptive Filters) *If the input date $\{h_i\}$ is exciting, all H^∞ optimal aposteriori adaptive filters that achieve $\gamma_{f,opt} = 1$ are given by*

$$\hat{z}_{j|j} = h_j\hat{w}_{|j} + (1+\mu h_j h_j^*)^{-\frac{1}{2}}\mathcal{S}_j\left((1+\mu h_j h_j^*)^{\frac{1}{2}}(d_j - h_j\hat{w}_{|j}), \ldots, (1+\mu h_0 h_0^*)^{\frac{1}{2}}(d_0 - h_0\hat{w}_{|0})\right) \tag{45}$$

where $\hat{w}_{|j}$ satisfies the recursion

$$\hat{w}_{|j+1} = \hat{w}_j + \frac{\mu h_{j+1}^*}{1+\mu h_{j+1}h_{j+1}^*}(d_{j+1} - h_{j+1}\hat{w}_{|j}) - \frac{\mu h_j^*}{1+\mu h_{j+1}h_{j+1}^*}(\hat{z}_{j|j} - h_j\hat{w}_{|j}), \quad \hat{w}_{|-1} \tag{46}$$

and $\mathcal{S}$ is any (possibly nonlinear) contractive causal mapping.

Proof: Simply restating the result of Theorem 3 for the special case $F_j = I$, $G_j = 0$, $H_j = h_j$ and $L_j = h_j$, and using the identity

$$I - h_j(P_j^{-1} + h_j^* h_j)^{-1}h_j^* = (I + h_j P_j h_j^*)^{-1},$$

along with the fact that for the H^∞-optimal a posteriori adaptive filters we have $\gamma_{f,opt} = 1$ and $P_i = \mu I$, yields the desired result.

■

We can now note the significance of some special choices for the causal contraction $\mathcal{S}$.

(i) $\mathcal{S} = 0$: This yields the normalized LMS algorithm.

(ii) $\mathcal{S} = I$: This yields

$$\hat{z}_{j|j} = h_j\hat{w}_{|j} + (1+\mu h_j h_j^*)^{-\frac{1}{2}}(1+\mu h_j h_j^*)^{\frac{1}{2}}(d_j - h_j\hat{w}_{|j}) = d_j,$$

which is the simple minded estimator of Section 6.1.

(iii) $\mathcal{S} = -I$: This yields

$$\hat{z}_{j|j} = h_j\hat{w}_{|j} - (1+\mu h_j h_j^*)^{-\frac{1}{2}}(1+\mu h_j h_j^*)^{\frac{1}{2}}(d_j - h_j\hat{w}_{|j}) = 2h_j\hat{w}_{|j} - d_j,$$

so that the recursion for $\hat{w}_{|j}$ becomes

$$\hat{w}_{|j+1} = \hat{w}_j + \frac{\mu h_{j+1}^*}{1+\mu h_{j+1}h_{j+1}^*}(d_{j+1} - h_{j+1}\hat{w}_{|j}) + \frac{\mu h_j^*}{1+\mu h_{j+1}h_{j+1}^*}(d_j - h_j\hat{w}_{|j}), \quad \hat{w}_{|-1}.$$

Theorem 8 (All H^∞ Apriori Adaptive Filters) *If the input data $\{h_i\}$ is exciting, and $0 < \mu < \inf_i \frac{1}{h_i h_i^*}$, then all H^∞ optimal apriori adaptive filters are given by*

$$\hat{z}_j = h_j\hat{w}_{|j-1} + (1-\mu h_j h_j^*)^{\frac{1}{2}}\mathcal{S}_j\left((1-\mu h_{j-1}h_{j-1}^*)^{\frac{1}{2}}(d_{j-1} - h_{j-1}\bar{w}_{|j-2}), \ldots, (1-\mu h_0 h_0^*)^{\frac{1}{2}}(d_0 - h_0\bar{w}_{|-1})\right), \tag{47}$$

where

$$\bar{w}_{|k-1} = \hat{w}_{|k-1} + \frac{\mu h_k^*}{-1+\mu h_k h_k^*}(\hat{z}_k - h_k\hat{w}_{|k-1}), \tag{48}$$

$\hat{w}_{|j}$ satisfies the recursion

$$\hat{w}_{|j} = \hat{w}_{|j-1} + \mu h_j^*(d_j - h_j\hat{w}_{|j-1}) - \mu h_j^*(\hat{z}_j - h_j\hat{w}_{|j-1}), \quad \hat{w}_{|-1} \tag{49}$$

and $\mathcal{S}$ is any (possibly nonlinear) contractive causal mapping.

Proof: Simply restating the result of Theorem 4 for the special case $F_j = I$, $G_j = 0$, $H_j = h_j$ and $L_j = h_j$, and using the fact that for the H^∞-optimal a priori filter we have $\gamma_{p,opt} = 1$, $P_i = \mu I$ and $\tilde{P}_i = \mu I - h_i^* h_i$, yields the desired result. Indeed equations (47), (48) and (49) are the corresponding specializations of equations (28), (29) and (30), respectively.

■

We once more note the consequences of some special choices of the causal contraction $\mathcal{S}$.

(i) $\mathcal{S} = 0$: This yields the LMS algorithm.

(ii) $\mathcal{S} = I$: This yields

$$\hat{z}_j = h_j\hat{w}_{|j-1} + (1-\mu h_j h_j^*)^{\frac{1}{2}}(1-\mu h_{j-1}h_{j-1}^*)^{\frac{1}{2}}(d_{j-1} - h_{j-1}\bar{w}_{|j-2}),$$

where $\bar{w}_{|j-2}$ and $\hat{w}_{|j-1}$ satisfy (48) and (49). The above filter is the a priori adaptive filter that corresponds to the simple minded estimator of Section 6.1. Note that in this case the filter is highly nontrivial.

(iii) $\mathcal{S} = -I$: This yields

$$\hat{z}_j = h_j \hat{w}_{|j-1} - (1 - \mu h_j h_j^*)^{\frac{1}{2}} (1 - \mu h_{j-1} h_{j-1}^*)^{\frac{1}{2}} (d_{j-1} - h_{j-1} \bar{w}_{|j-2}).$$

Note that it does not seem possible to obtain a simplistic a priori estimator that achieves optimal performance.

8 Risk-Sensitive Optimality

In this section we focus on a certain property of the central H^∞ filters, namely the fact that they are risk-sensitive optimal filters. This will give further insight into the LMS and normalized LMS algorithms, and in particular will provide a stochastic interpretation in the special case of disturbances that are independent Gaussian random variables.

The risk-sensitive (or exponential cost) criterion was introduced in [14] and further studied in [15, 16, 17]. We begin with a brief introduction to the risk-sensitive criterion. For much more on this subject consult [16].

8.1 The Exponential Cost Function

Although it is straightforward to consider the risk-sensitive criterion in the full generality of the state-space model of Section 4, here we only deal with the special case of our interest. To this end, consider the state-space model corresponding to the adaptive filtering problem we have been studying:

$$\left\{ \begin{array}{rcl} x_{i+1} & = & x_i \\ d_i & = & h_i x_i + v_i \end{array} \right. , \quad x_0 = w \tag{50}$$

where we now assume that w and the $\{v_i\}$ are independent Gaussian random variables with means $\hat{w}_{|-1}$ and zero and covariances Π_0 and I, respectively. As before, we are interested in the filtered and predicted estimates $\hat{z}_{i|i} = \mathcal{F}_f(d_0, d_1, \ldots, d_i)$ and $\hat{z}_i = \mathcal{F}_p(d_0, d_1, \ldots, d_{i-1})$ of the uncorrupted output $z_i = h_i x_i$. The corresponding filtered and predicted errors are given by $e_{f,i} = \hat{z}_{i|i} - z_i$ and $e_{p,i} = \hat{z}_i - z_i$. The conventional Kalman filter is an estimator that performs the following minimization (see *e.g.* [25, 26]):

$$\min_{\{\hat{z}_j\}} \left[E \sum_{j=0}^{i} e_{p,j}^* e_{p,j} \right], \tag{51}$$

where the expectation is taken over the Gaussian random variables w and $\{v_j\}_{j=0}^{\infty}$ whose joint conditional distribution is given by:

$$p(w, v_0, \ldots, v_i | d_0, \ldots, d_i) \propto$$

$$exp\left[-\frac{1}{2}\left((w - \hat{w}_{|-1})^* \Pi_0^{-1} (w - \hat{w}_{|-1}) + \sum_{j=0}^{i} (d_j - h_j x_j)^* (d_j - h_j x_j) \right) \right],$$

and where the symbol $\propto$ stands for 'proportional to'. In the terminology of [16], the filter that minimizes (51) is known as the *risk-neutral* filter.

An alternative criterion that is risk-sensitive has been extensively studied in [14] - [17] and corresponds to the following minimization problem

$$\min_{\{\hat{z}_{j|j}\}} \mu_{f,i}(\theta) = \min_{\{\hat{z}_{j|j}\}} \left(-\frac{2}{\theta} log \left[Eexp(-\frac{\theta}{2} \mathbf{C}_{f,i}) \right] \right), \tag{52a}$$

or

$$\min_{\{\hat{z}_j\}} \mu_{p,i}(\theta) = \min_{\{\hat{z}_j\}} \left(-\frac{2}{\theta} log \left[Eexp(-\frac{\theta}{2} \mathbf{C}_{p,i}) \right] \right), \tag{52b}$$

where $\mathbf{C}_{f,i} = \sum_{j=0}^{i} e_{f,i}^* e_{f,i}$ and $\mathbf{C}_{p,i} = \sum_{j=0}^{i} e_{p,i}^* e_{p,i}$. The criteria in (52a) and (52b) are known as the a posteriori and a priori *exponential cost functions*, and any filters that minimize $\mu_{f,i}(\theta)$ and $\mu_{p,i}(\theta)$ are referred to as a posteriori and a priori risk-sensitive filters, respectively. The scalar parameter θ is correspondingly called the *risk-sensitivity* parameter. Some intuition concerning the nature of this modified criterion is obtained by expanding $\mu_i(\theta)$ (where we have dropped the subscripts f and p since the argument follows for both filtered and predicted estimates) in terms of θ and writing,

$$\mu_i(\theta) = E(\mathbf{C}_i) - \frac{\theta}{4} Var(\mathbf{C}_i) + O(\theta^2).$$

The above equation shows that for $\theta = 0$, we have the risk-neutral case (*i.e.,* the conventional Kalman filter). When $\theta > 0$, we seek to maximize $Eexp(-\frac{\theta}{2}\mathbf{C}_i)$, which is convex and decreasing in $\mathbf{C}_i$. Such a criterion is termed *risk-seeking* (or optimistic) since larger weights are on small values of $\mathbf{C}_i$, and hence we are more concerned with the frequent occurrence of moderate values of $\mathbf{C}_i$ than with the occasional large values. When $\theta < 0$, we seek to minimize $Eexp(-\frac{\theta}{2}\mathbf{C}_i)$, which is convex and increasing in $\mathbf{C}_i$. 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.

The relationship between the risk-sensitive criterion and the H^∞ criterion was first noted in [27] and has been further discussed in [16, 19]. It may be formally stated as follows: *In the risk-averse case $\theta < 0$, the risk-sensitive optimal filter with parameter θ is given by the central H^∞ filter with level $\gamma = -\theta^{-\frac{1}{2}}$.* In particular, there is a certain *smallest* value of the risk-sensitivity parameter $\bar{\theta}$, after which the minimizing property of $\mu_i(\theta)$ breaks down, and it is this value that yields the optimal central H^∞ filter with $\gamma_{opt} = -\bar{\theta}^{-1/2}$.

8.2 Risk-sensitive Adaptive Filtering

Using the discussion of Section 8.1, we are now in a position to state the risk-sensitive results for LMS and normalized LMS.

Theorem 9 (Normalized LMS and Risk-sensitivity) *Consider the state-space model (50) where the w and $\{v_j\}$ are independent Gaussian random variables with means $\hat{w}_{|-1}$ and 0, and variances μI and I, respectively. The solution to the following minimization problem*

$$\min_{\{\hat{z}_{j|j}\}} \mu_f(\theta) = \min_{\{\hat{z}_{j|j}\}} \left(2log\left[Eexp(\frac{1}{2}\mathbf{C}_f),\right]\right) \tag{53}$$

where $\mathbf{C}_f = \sum_{j=0}^{\infty} e_{f,i}^* e_{f,i}$, and the expectation is taken over w and $\{v_j\}$ subject to observing $\{d_0, d_1, \ldots, d_i\}$, is given by the normalized LMS algorithm

$$\hat{z}_{i|i} = h_i \hat{w}_{|i},$$

and

$$\hat{w}_{|i+1} = \hat{w}_{|i} + \frac{\mu h_{i+1}^*}{1 + \mu h_{i+1} h_{i+1}^*}(d_{i+1} - h_{i+1}\hat{w}_{|i}) \quad , \quad \hat{w}_{|-1}. \tag{54}$$

Theorem 10 (LMS and Risk-sensitivity) *Consider the state-space model (50) where the w and $\{v_j\}$ are independent Gaussian random variables with means $\hat{w}_{|-1}$ and 0, and variances μI and I, respectively. Suppose moreover, that the $\{h_i\}$ are exciting, and that*

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

Then the solution to the following minimization problem

$$\min_{\{\hat{z}_j\}} \mu_p(\theta) = \min_{\{\hat{z}_j\}} \left(2log \left[Eexp(\frac{1}{2}\mathbf{C}_p), \right] \right) \tag{55}$$

where $\mathbf{C}_p = \sum_{j=0}^{\infty} e_{p,i}^* e_{p,i}$, 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

$$\hat{z}_i = h_i \hat{w}_{i-1},$$

and

$$\hat{w}_{|i} = \hat{w}_{|i-1} + \mu h_i^* (d_i - h_i \hat{w}_{|i-1}) \quad , \quad \hat{w}_{|-1}. \tag{56}$$

Before closing this section we should remark that the central H^∞ filters possess other properties in addition to the one described above. In the game theoretic formulation of H^∞ estimation, the central filter corresponds to the *solution* of the game [28]. Morever, among all H^∞ estimators that achieve a certain level γ, the central solution can be shown to be the maximum entropy [21] solution. However, we shall not pursue these directions here.

9 Further Remarks

In addition to yielding a new interpretation for the LMS algorithm and providing it with a rigorous basis, the results described in this paper have lent themselves to various generalizations and have allowed the authors to obtain several new results. We close this paper by listing some of these ideas and results here. We should also mention that we believe the framework presented in this paper provides a new way of looking at adaptive algorithms and should be worthy of further scrutiny.

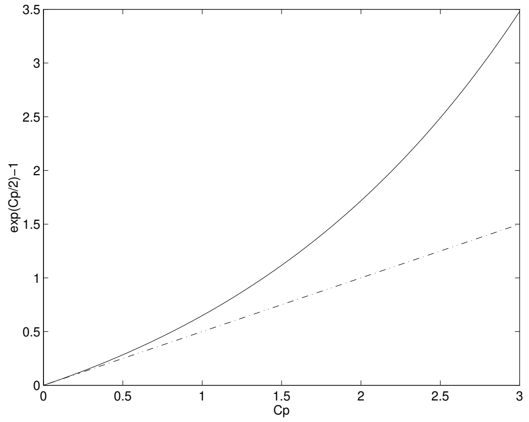

Figure 6: *The criterion (55) is termed risk averse (or pessimistic) since the cost function* $exp(\mathbf{C}_p/2)$ *is very large for large values of* $\mathbf{C}_p$. *Hence we are more concerned with the occasional occurence of large values of* $\mathbf{C}_p$ *than with the frequent occurrence of moderate ones. This fact corresponds well with the intuition gained from the* H^∞ *optimality of the LMS algorithm. We have also plotted* $\mathbf{C}_p/2$ *(the dashed line) to compare the two cost functions, since the RLS algorithm minimizes the expected value of* $\mathbf{C}_p/2$.

LMS with Time-Varying Learning Rate

In many applications one uses the LMS algorithm with time-varying stepsize (or learning rate), viz.,

$$\hat{w}_{|i} = \hat{w}_{|i-1} + \mu_i h_i^*(d_i - h_i\hat{w}_{|i-1}), \quad \hat{w}_{|-1}. \tag{57}$$

In this case, it is straightforward to show that if the vectors $\{\mu_i^{1/2}h_i\}$ are exciting, and if $\mu_i h_i h_i^* \leq 1$ for all i, then the LMS algorithm with time-varying stepsize solves the following minimax problem:

$$\inf_{\mathcal{F}} \sup_{w,v\in h_2} \frac{\sum_{j=0}^{\infty}\mu_j|e_{p,j}|^2}{|w-\hat{w}_{|-1}|^2 + \sum_{j=0}^{\infty}\mu_j|v_j|^2} = 1. \tag{58}$$

H^∞ Adaptive Filtering

In this paper we have shown that if adaptive filtering for *output prediction error* is considered then the central H^∞-optimal adaptive filter is LMS. It is also possible to consider prediction of the filter weight vector itself, and for the purpose of coping with time-variations, to consider exponentially weighted, finite-memory and time-varying adaptive filtering. This results in some new adaptive filtering algorithms that may be useful in uncertain and non-stationary environments (see [29]).

H^∞ Norm Bounds for the RLS Algorithm

In order to compare the robustness of H^2-optimal algorithms (such as RLS) with H^∞-optimal algorithms (such as LMS) it it useful to obtain H^∞ norm bounds for these algorithms. This has been done for the RLS algorithm in [20], where it is shown that unlike LMS, the H^∞ norm of the RLS algorithm depends on the input data $\{h_i\}$ and, roughly speaking, grows linearly in the parameter μ.

A Time-Domain Feedback Analysis

Using some of the ideas presented here, a time-domain feedback analysis of recursive adaptive schemes, including gradient-based and Gauss-Newton filters has been developed [30, 31], for both the FIR and IIR contexts. The analysis highlights an intrinsic feedback structure in terms of a feedforward lossless or contractive map and a feedback memoryless or dynamic map. The structure lends itself to analysis via energy conservation arguments and via standard tools in system theory such as the small gain theorem [32]. It further suggests choices for the adaptation gains (or step-sizes) in order to enforce a robust performance in the presence of disturbances (along the lines of H^∞theory), as well as improve the convergence speed of the adaptive algorithms.

Nonlinear Problems

The results presented in this paper are for linear adaptive filters and can be somewhat generalized to nonlinear adaptive filters (such as neural networks) if one linearizes these nonlinear models around some suitable point. Using this approach it can be shown (see [34]) that, for nonlinear problems, instantaneous-gradient-based algorithms (such as backpropagation [33]) are *locally* H^∞-optimal. This means that if the initial estimate of the weight vector is close enough to its true value, and if the disturbances are small enough, then the maximum energy gain from the disturbances to the output prediction errors is arbitrarily close to one. *Global* H^∞-optimal filters can also be found in the nonlinear case, but they have the drawback of being infinite-dimensional [35].

10 Conclusion

We have demonstrated that the LMS algorithm is H^∞ optimal. This result solves a long standing issue of finding a rigorous basis for the LMS algorithm, and also confirms its robustness. We find it quite interesting that despite the fact that there has only been recent interest in the field of H^∞ estimation, there has existed an H^∞ optimal estimation algorithm that has been widely used in practice for the past three decades.

Acknowledgement

The first author would like to thank Professor L. Ljung for contributing to the discussion in Section 6.1.

References

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

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

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

[4] A. H. Sayed and T. Kailath. A state-space approach to adaptive RLS filtering. *IEEE Signal Processing Magazine*, July 1994, pp. 18-60, Vol 11, no 3.

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

[6] H. Kwakernaak. A polynomial approach to mimimax frequency domain optimization of multivariable feedback systems. *Int. J. of Control*, 44:117–156, 1986.

[7] J. C. Doyle, K. Glover, P. Khargonekar, and B. Francis. State-space solutions to standard H_2 and H_∞ control problems. *IEEE Transactions on Automatic Control*, 34(8):831–847, August 1989.

[8] 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.

[9] T. Basar. Optimum performance levels for minimax filters, predictors and smoothers. *Systems and Control Letters*, 16:309–317, 1991.

[10] D. J. Limebeer and U. Shaked. New results in H^∞-filtering. In *Proc. Int. Symp. on MTNS*, pages 317–322, June 1991.

[11] U. Shaked and Y. Theodor. $H^\infty-$optimal estimation: A tutorial. In *Proc. IEEE Conference on Decision and Control*, pages 2278–2286, Tucson, AZ, Dec. 1992.

[12] M. J. Grimble. Polynomial matrix solution of the H^∞ filtering problem and the relationship to Riccati equation state-space results. *IEEE Trans. on Signal Processing*, 41(1):67–81, January 1993.

[13] M. Green and D.J.N. Limebeer. *Linear Robust Control.* Prentice Hall, Englewood Cliffs NJ, 1995.

[14] D. H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to determenistic games. *IEEE Trans. Automatic Control*, 18(2), April 1973.

[15] J. Speyer, J. Deyst, and D. H. Jacobson. Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria. *IEEE Trans. Automatic Control*, 19:358–366, August 1974.

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

[17] J.L. Speyer, C. Fan, and R. N. Banavar. Optimal stochastic estimation with exponential cost criteria. In *Proc. IEEE Conference on Decision and Control*, pages 2293–2298, Tucson, Arizona, December 1992.

[18] B. Hassibi, A. H. Sayed, and T. Kailath. Recursive linear estimation in Krein spaces - part I: Theory. In the *IEEE Transactions on Automatic Control*, Jan 1996.

[19] B. Hassibi, A. H. Sayed, and T. Kailath. Recursive linear estimation in Krein spaces - Part II: Applications. In the *IEEE Transactions on Automatic Control*, Jan 1996.

[20] B. Hassibi and T. Kailath. H^∞ Bounds for the recursive-least-squares algorithm. In *Proc. IEEE Conference on Decision and Control*, Orlando, FL, Dec. 1994, pp. 3927-3929.

[21] 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.

[22] B. Hassibi, B. Halder and T. Kailath. Mixed H^2/H^∞ estimation: preliminary analytic characterization and a numerical solution. To appear in the *13th World Congress International Federation of Automatic Control*, Jun 30 - Jul 5, 1996, San Francisco, CA.

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

[24] B. Widrow, et al. Stationary and nonstationary learning characteristics of the LMS adaptive filter. *Proceedings IEEE*, 64(8):1151–1162, August 1976.

[25] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, New York, 1970.

[26] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall Inc., NJ, 1979.

[27] K. Glover and J. C. Doyle. State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to risk sensitivity. *System and Control Letters*, 11:167–172, 1988.

[28] T. Basar and P. Bernhard. *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach* Birkhauser, Boston, 1991.

[29] B. Hassibi and T. Kailath. H^∞ Adaptive Filtering, Proc. *1995 IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, MI, May 1995.

[30] A.H. Sayed and M. Rupp. A time-domain feedback analysis of adaptive gradient algorithms via the small gain theorem, *Proc. SPIE Conference*, vol 2563, pp. 458-469, San Diego CA, July 1995.

[31] A.H. Sayed and M. Rupp. A class of adaptive nonlinear $H^\infty-$filters with guaranteed l_2-stability, In *Proc. 3rd IFAC Symposium on Nonlinear Control Systems Design*, vol. 2, pp. 455–460, June 1995.

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

[33] 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.

[34] B. Hassibi, A.H. Sayed and T. Kailath, H^∞ Optimality Criteria for LMS and Backpropagation, in *Advances in Neural Information Processing Systems*, Vol 6, J.D. Cowan, G. Tesauro and J. Alspector, Eds., pp. 351-359, Morgan-Kaufmann, Apr 1994.

[35] B. Hassibi and T. Kailath, H^∞ Optimal Training Algorithms and their Relation to Backpropagation, In *Advances in Neural Information Processing Systems*, Vol 7, G. Tesauro, D.S. Touretzky and T.K. Leen , Eds., pp. 191-199, MIT Press, Apr 1995.

A A First Principles Proof of the H^∞ Optimality of LMS

In this appendix we shall outline a first principles proof of the H^∞ optimality of the LMS and normalized LMS algorithms that does not require the results of Theorems 1 and 2 on H^∞ filtering. The proofs rely on some easily verified inequalities. We begin with normalized LMS. (See also the last section in [4] and [30].)

A.1 The Normalized LMS Algorithm

Recall that in Sec. 5.1, after the statement of Theorem 5, we constructed a disturbance signal such that for any $\epsilon > 0$,

$$\frac{\|e_f\|^2}{\mu^{-1}|w-\hat{w}_{|-1}|^2+\|v\|^2} \geq 1-\epsilon.$$

Since this was just one special disturbance signal, we conclude that if the input vectors are exciting, we have

$$\sup_{w,v\in h_2} \frac{\|e_f\|^2}{\mu^{-1}|w-\hat{w}_{|-1}|^2+\|v\|^2} \geq 1. \tag{1}$$

We shall now show that the normalized LMS algorithm achieves one in the above inequality. This, of course, also shows that $\gamma_{f,opt} = 1$. To this end, note that the normalized LMS algorithm

$$\hat{w}_{|j} = \hat{w}_{|j-1} + \frac{h_j^*}{\mu^{-1}+h_jh_j^*}(d_j - h_j\hat{w}_{|j-1}),$$

can, after some rearrangement, be written as

$$\hat{w}_{|j-1} = \hat{w}_{|j} - \mu h_i^*(d_j - h_j\hat{w}_{|j}).$$

If we now define $\tilde{w}_{|j} = w - \hat{w}_{|j}$, the above expression allows us to write

$$\mu^{-1/2}\left[\tilde{w}_{|j-1}\right] = \mu^{-1/2}\left[\tilde{w}_{|j} + \mu h_i^*(d_j - h_j\hat{w}_{|j})\right]. \tag{2}$$

[The reason for multiplying both sides by $\mu^{-1/2}$ will become clear in a moment.] On the other hand, we may write $v_j = d_j - h_j w$ as

$$v_j = (d_j - h_j \hat{w}_{|j}) - h_j \tilde{w}_{|j}. \tag{3}$$

Squaring both sides of (2) and (3) and adding the results yields

$$\mu^{-1}|\tilde{w}_{|j-1}|^2 + |v_j|^2 = \mu^{-1}|\tilde{w}_{|j}|^2 + |h_j \tilde{w}_{|j}|^2 + (1 + \mu h_j h_j^*)(d_j - h_j \hat{w}_{|j})^2. \tag{4}$$

Now since the third term on the RHS of the above expression is positive, and since $h_j \tilde{w}_{|j} = e_{f,j}$, we may write

$$\mu^{-1}|\tilde{w}_{|j-1}|^2 + |v_j|^2 \geq \mu^{-1}|\tilde{w}_{|j}|^2 + |e_{f,j}|^2. \tag{5}$$

If we now add all inequalities of the form (5) from time $j = 0$ to time $j = i$, we have

$$\mu^{-1}|w - \hat{w}_{|-1}|^2 + \sum_{j=0}^{i} |v_j|^2 \geq \mu^{-1}|\tilde{w}_{|i}|^2 + \sum_{j=0}^{i} |e_{f,j}|^2 \geq \sum_{j=0}^{i} |e_{f,j}|^2, \tag{6}$$

which in turn implies

$$\frac{\sum_{j=0}^{i} |e_{f,j}|^2}{\mu^{-1}|w - \hat{w}_{|-1}|^2 + \sum_{j=0}^{i} |v_j|^2} \leq 1. \tag{7}$$

Thus, for normalized LMS, in the limit as $i \to \infty$ we have

$$\sup_{w,v \in h_2} \frac{\sum_{j=0}^{\infty} |e_{f,j}|^2}{\mu^{-1}|w - \hat{w}_{|-1}|^2 + \sum_{j=0}^{\infty} |v_j|^2} = \frac{\|e_f\|^2}{\mu^{-1}|w - \hat{w}_{|-1}|^2 + \|v\|^2} = 1, \tag{8}$$

which is the desired result.

A.2 The LMS Algorithm

The proof for the LMS algorithm follows the exact same lines as the one above. Eq. (2) is now replaced by

$$\mu^{-1/2}\left[\tilde{w}_{|j}\right] = \mu^{-1/2}\left[\tilde{w}_{|j-1} - \mu h_i^*(d_j - h_j \hat{w}_{|j-1})\right], \tag{9}$$

and (3) by

$$v_j = (d_j - h_j \hat{w}_{|j-1}) - h_j \tilde{w}_{|j-1}. \tag{10}$$

This time we square both sides of (9) and (10) and subtract the results to obtain

$$\mu^{-1}|\tilde{w}_{|j}|^2 - |v_j|^2 = \mu^{-1}|\tilde{w}_{|j-1}|^2 - |h_j \tilde{w}_{|j-1}|^2 - (1 - \mu h_j h_j^*)(d_j - h_j \hat{w}_{|j-1})^2. \tag{11}$$

Now since we have the bound $\mu \leq \frac{1}{h_j h_j^*}$, the third term on the RHS is negative, and we can write

$$\mu^{-1}|\tilde{w}_{|j-1}|^2 + |v_j|^2 \geq \mu^{-1}|\tilde{w}_{|j}|^2 + |\underbrace{h_j \tilde{w}_{|j-1}}_{e_{p,j}}|^2. \tag{12}$$

The remainder of the proof is now identical to the normalized LMS case.