In *Adaptive Control, Filtering, and Signal Processing*, IMA Volumes in Mathematics and Its Applications, vol. 74, K. J. Åström, G. C. Goodwin, and P. R. Kumar, eds., pp. 65–88, Springer-Verlag, NY, 1995.

LMS is $\mathcal{H}^\infty$-Optimal *

Babak Hassibi† Ali H. Sayed and Thomas Kailath

Abstract

We show that the celebrated LMS (Least-Mean Squares) adaptive algorithm is H^∞ optimal. In other words, the LMS algorithm, which has long been regarded as an approximate least-mean squares solution, is in fact an exact minimizer of a certain so-called H^∞ error norm. In particular, the LMS minimizes the energy gain from the disturbances to the *predicted* errors, while the so-called normalized LMS minimizes the 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 (in the sense of Whittle). 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.

*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

The LMS algorithm was originally conceived as an approximate recursive procedure that solves the following adaptive problem [1, 2]: given a sequence of $1 \times M$ input row vectors $\{h_i\}$, and a corresponding sequence of desired responses $\{d(i)\}$, find an estimate of an $M \times 1$ column vector of weights w such that the sum squared error $\sum_{i=0}^{N} |d(i) - h_i w|^2$ is minimized. The LMS solution recursively updates estimates of the weight vector along the direction of the instantaneous gradient of the squared error. The introduction of the LMS 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. Despite the name, however, we should note that the LMS algorithm does not minimize the sum of squared errors, and has long been thought to be an approximate minimizing solution.

Algorithms that exactly minimize the sum of squared errors, for every value of N are known: they are the well-known recursive least squares (RLS) algorithms (see, e.g., [3]). They have better convergence properties, but are computationally more complex, and are less robust than the simple LMS algorithm. For example, it has been observed that the LMS has better tracking capabilities than the RLS algorithm in the presence of nonstationary inputs [3]. We show here 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 2, here we note only that the H^∞ criterion was introduced to address the fact that in many applications one is often faced with model uncertainties and lack of statistical information on the exogeneous signals. The great recent interest in H^∞ filtering may be seen from [4, 5, 6, 7, 8, 9, 10, 11] and the many other references therein.

In this paper, we shall use some of the well known results in H^∞ estimation theory in order to show that the LMS algorithm is the so-called central *apriori* H^∞-optimal filter, while the so-called normalized LMS algorithm is the central *aposteriori* H^∞-optimal filter. This provides LMS with a rigorous basis and furnishes a minimization criterion that has long been missing. Morever, since LMS and normalized LMS are shown to be *central* filters they also minimize a certain exponential cost function, and are thus risk-sensitive optimal [16].

The remainder of the paper is organized as follows. In Section 2 we review the H^∞ estimation problem as one that minimizes the energy gain from the disturbances to the estimation error. We consider the aposteriori and apriori cases which correspond to filtered and predicted estimation errors, respectively. Section 3 gives the expressions for the H^∞ aposteriori and apriori filters, as well as their full parametrization, since such filters are not unique. In Section 4, we formulate the H^∞ adaptive filtering problem. Section 5 shows that the normalized LMS algorithm is the central aposteriori H^∞ optimal adaptive filter, and that if the learning rate is chosen appropriately, LMS is the central apriori H^∞ optimal adaptive filter. We then consider a simple example that demonstrates the robustness of LMS compared to RLS, and in Section 5.4 present a discussion on the merit of the different H^∞ optimal algorithms. With this in mind, we develop the full parametrization of all H^∞ optimal adaptive filters in Section 6, and in Section 7 show that LMS and normalized LMS have the additional property of being risk-sensitive optimal. This provides LMS and normalized LMS with a stochastic interpretation in the special case of disturbances that are white Gaussian random variables. Section 8 offers a very brief summary.

We find it 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.

2 The H^∞ Problem

We first give a brief review of some of the results in H^∞ estimation theory using the notation of the companion papers [12, 13]. The reader is also referred to [4, 5, 6, 7, 8, 9, 10, 11] and the references therein for earlier results and alternative approaches.

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 H^∞ Problem

We now consider a state-space model of the form

$$\begin{aligned} x_{i+1} &= F_i x_i + G_i u_i \quad , \quad x_0 \\ y_i &= H_i x_i + v_i \end{aligned} \tag{1}$$

where 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, 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 the estimate of z_i given observations $\{y_j\}$ from time 0 up to and including time i, and $\hat{z}_i = \mathcal{F}_p(y_0, y_1, \ldots, y_{i-1})$ denote the estimate of z_i given observations $\{y_j\}$ from time 0 to time $i-1$. We then have the following two estimation errors: the *filtered* error

$$e_{f,i} = \hat{z}_{i|i} - L_i x_i, \tag{2}$$

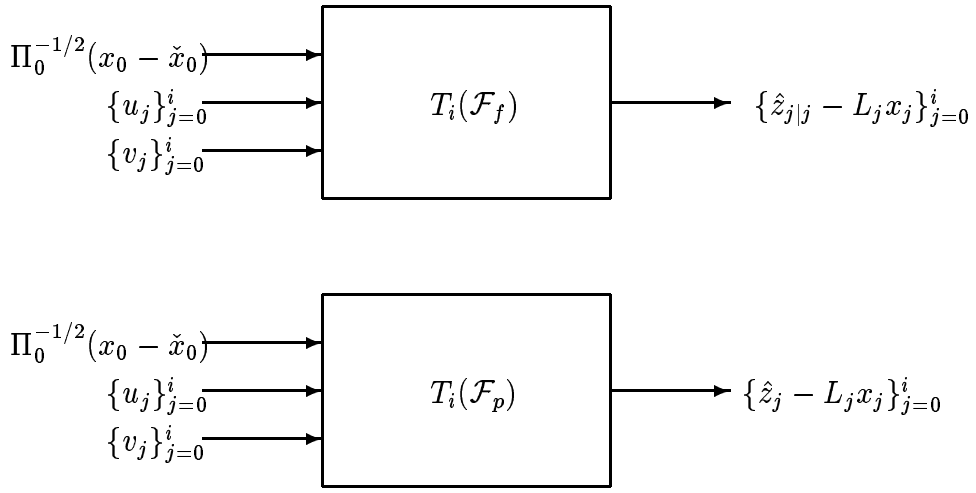

Figure 1: Transfer matrices from disturbances to filtered and predicted estimation error

and the *predicted* error

$$e_{p,i} = \hat{z}_i - L_i x_i. \tag{3}$$

Let T_f (T_p) denote the transfer operator that maps the unknown disturbances $\{x_0 - \hat{x}_0, u_i, v_i\}$ (where $\hat{x}_0$ denotes an initial guess of x_0) to the filtered (predicted) error $e_{f,i}$ ($e_{p,i}$). See Figure 1. The H^∞ estimation problem can now be stated as follows.

Problem 1 (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 $\| T_f \|_\infty$ and $\| T_p \|_\infty$, and obtain the resulting*

$$\gamma_{f,o}^2 = \inf_{\mathcal{F}_f} \| T_f \|_\infty^2 = \inf_{\mathcal{F}_f} \sup_{x_0, u \in h_2, v \in h_2} \frac{\| e_f \|_2^2}{(x_0 - \hat{x}_0)^* \Pi_0^{-1} (x_0 - \hat{x}_0) + \| u \|_2^2 + \| v \|_2^2} \tag{4}$$

and

$$\gamma_{p,o}^2 = \inf_{\mathcal{F}_p} \| T_p \|_\infty^2 = \inf_{\mathcal{F}_p} \sup_{x_0, u \in h_2, v \in h_2} \frac{\| e_p \|_2^2}{(x_0 - \hat{x}_0)^* \Pi_0^{-1} (x_0 - \hat{x}_0) + \| u \|_2^2 + \| v \|_2^2} \tag{5}$$

where Π_0 is a positive definite matrix that reflects apriori knowledge as to how close x_0 is to the initial guess $\hat{x}_0$.

Note that the infimum in (5) is taken over all *strictly* causal estimators $\mathcal{F}_p$, whereas in (4) 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 to be studied), and a simpler problem results if one relaxes the minimization condition and settles for a suboptimal solution.

Problem 2 (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 $\| T_f \|_\infty \leq \gamma_f$ and $\| T_p \|_\infty \leq \gamma_p$. This clearly requires checking whether $\gamma_f \geq \gamma_{f,o}$ and $\gamma_p \geq \gamma_{p,o}$.*

To guarantee $\| T_f \|_\infty \leq \gamma_f$ we shall proceed as follows. Let $T_{f,i}$ be the transfer operator that maps the disturbances $\left\{x_0 - \hat{x}_0, \{u_j\}_{j=0}^i, \{v_j\}_{j=0}^i\right\}$ to the filtered errors $\left\{\{e_{f,j}\}_{j=0}^i\right\}$. We shall find a γ_f that ensures $\| T_{f,i} \|_\infty < \gamma_f$ for all i. Likewise we shall find a γ_p that ensures for $\| T_{p,i} \|_\infty < \gamma_p$ for all i.

3 The H^∞ Filters

We now briefly review some of the results on H^∞ filters using the notation of [12, 13].

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

$$P_j^{-1} + H_j^* H_j - \gamma^{-2} L_j^* L_j > 0, \qquad j = 0, \ldots, i \tag{6}$$

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

$$\begin{aligned} P_{j+1} \;=\; & F_j P_j F_j^* + G_j G_j^* - \\ & F_j P_j \begin{bmatrix} L_j^* & H_j^* \end{bmatrix} \left\{ \begin{bmatrix} -\gamma^2 I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} L_j \\ H_j \end{bmatrix} P_j \begin{bmatrix} L_j^* & H_j^* \end{bmatrix} \right\}^{-1} \begin{bmatrix} L_j \\ H_j \end{bmatrix} P_j F_j^* \end{aligned} \tag{7}$$

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

$$\hat{x}_{j+1|j+1} = F_j \hat{x}_{j|j} + K_{f,j}(y_{j+1} - H_{j+1} F_j \hat{x}_{j|j}) \quad , \quad \hat{x}_{-1|-1} \tag{8}$$

and

$$K_{f,j} = P_{j+1} H_{j+1}^* (I + H_{j+1} P_{j+1} H_{j+1}^*)^{-1} \tag{9}$$

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

$$\tilde{P}_j^{-1} = P_j^{-1} - \gamma^{-2} L_j^* L_j > 0, , \qquad j = 0, \ldots, i \tag{10}$$

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

$$\hat{z}_j = L_j \hat{x}_j \tag{11}$$

$$\hat{x}_{j+1} = F_j \hat{x}_j + K_{p,j}(y_j - H_j \hat{x}_j) \quad , \quad \hat{x}_0 \tag{12}$$

where

$$K_{p,j} = F_j \tilde{P}_j H_j^* (I + H_j \tilde{P}_j H_j^*)^{-1} \tag{13}$$

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

$$\begin{aligned} \hat{x}_{j+1} &= F_j\hat{x}_j + F_jP_jH_j^*(I + H_jP_jH_j^*)^{-1}(y_j - H_j\hat{x}_j) \\ P_{j+1} &= F_jP_jF_j^* + G_jG_j^* - F_jP_j(I + H_jP_jH_j^*)^{-1}P_jF_j^* \end{aligned} \tag{14}$$

and that the only difference is that the P_j of equation (9), and $\tilde{P}_j$ of equation (13), satisfy Riccati recursions that differ with (14). However, as $\gamma \to \infty$, the Riccati recursion (7) collapses to the Kalman filter recursion (14). 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 (7), 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_ix_i.

Note also that condition (10) is more stringent than condition (6), indicating that the existence of an apriori filter of level γ implies the existence of an aposteriori 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 [13]).

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

$$\begin{aligned} \hat{z}_{i|i} &= L_i\hat{x}_{i|i} + [I - L_i(P_i^{-1} + H_i^*H_i)^{-1}L_i^*]^{\frac{1}{2}} \\ &\quad \mathcal{S}_i\left((I + H_iP_iH_i^*)^{\frac{1}{2}}(y_i - H_i\hat{x}_{i|i}), \ldots, (I + H_0P_0H_0^*)^{\frac{1}{2}}(y_0 - H_0\hat{x}_{0|0})\right) \end{aligned} \tag{15}$$

where $\hat{x}_{i|i}$ is given by Theorem 1, and

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

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

$$\sum_{j=0}^{i} |\mathcal{S}_j(a_j, \ldots, a_0)|^2 \leq \sum_{j=0}^{i} |a_j|^2$$

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

$$\begin{aligned} \hat{z}_i &= L_i\hat{x}_i + [I - L_iP_iL_i^*]^{\frac{1}{2}} \\ &\quad \mathcal{S}_i\left((I + H_{i-1}\tilde{P}_{i-1}H_{i-1}^*)^{-\frac{1}{2}}(y_{i-1} - H_{i-1}\hat{x}_{i-1}), \ldots, (I + H_0\tilde{P}_0H_0^*)^{-\frac{1}{2}}(y_0 - H_0\hat{x}_0)\right) \end{aligned} \tag{16}$$

where $\hat{x}_i$ and $\tilde{P}_i$ are 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 are 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 $\mathcal{S} = 0$.

4 Formulation of the H^∞ Adaptive Problem

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

$$d_i = h_i w + v_i \tag{17}$$

where $h_i = \begin{bmatrix} h_1(i) & h_2(i) & \dots & h_M(i) \end{bmatrix}$ is a known $1 \times M$ vector whose elements are the inputs from M input channels ($h_k(i)$ denotes the input at time i to the kth channel), $w = \begin{bmatrix} w_1 & w_2 & \dots & w_M \end{bmatrix}^T$ is an unknown $M \times 1$ weight vector, 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, normal distributed, etc.).

Note that equation (17) can be restated in the following state-space form:

$$\begin{aligned} x_{i+1} &= x_i \quad , \qquad x_0 = w \\ d_i &= h_i x_i + v_i \end{aligned} \tag{18}$$

This is a relevant step since it reduces the adaptive filtering problem to an equivalent state-space estimation problem. For example, the RLS algorithm follows if one applies the Kalman filter to (18). Here we shall show that the LMS and normalized LMS algorithms follow from applying the H^∞ theory to (18).

Consider the uncorrupted output $z_i = h_i x_i$ of (18). Let $\hat{z}_{i|i} = \mathcal{F}_f(d_0, d_1, \dots, d_i)$ denote the estimate of z_i using the noisy measurements $\{d_j\}$ and the input vectors $\{h_j\}$ from time 0 up to and including i. Likewise, let $\hat{z}_i = \mathcal{F}_p(d_0, d_1, \dots, d_{i-1})$ denote the estimate of z_i using the noisy measurements $\{d_j\}$ and the input vectors $\{h_j\}$ from time 0 to time $i-1$. As before, we have the *filtered* error

$$e_{f,i} = \hat{z}_{i|i} - h_i x_i, \tag{19}$$

and the *predicted* error

$$e_{p,i} = \hat{z}_i - h_i x_i. \tag{20}$$

Let T_f (T_p) denote the transfer operator that maps the unknown disturbances $\{w - \hat{w}_{|-1}, v_i\}$ (where $\hat{w}_{|-1}$ is an initial guess of w) to the filtered (predicted) error $e_{f,i}$ ($e_{p,i}$). The H^∞ adaptive filtering problem can then be stated as follows.

Problem 3 (Optimal H^∞ Adaptive Problem) *Find H^∞-optimal estimation strategies $\hat{z}_{i|i} = \mathcal{F}_f(d_0, d_1, \dots, d_i)$ and $\hat{z}_i = \mathcal{F}_p(d_0, d_1, \dots, d_{i-1})$ that respectively minimize $\| T_f \|_\infty$ and $\| T_p \|_\infty$, and obtain the resulting*

$$\gamma_{f,o}^2 = \inf_{\mathcal{F}_f} \| T_f \|_\infty^2 = \inf_{\mathcal{F}_f} \sup_{w,v \in h_2} \frac{\| e_f \|_2^2}{(w - \hat{w}_{|-1})^* \Pi_0^{-1} (w - \hat{w}_{|-1}) + \| v \|_2^2} \tag{21}$$

and

$$\gamma_{p,o}^2 = \inf_{\mathcal{F}_p} \| T_p \|_\infty^2 = \inf_{\mathcal{F}_p} \sup_{w,v \in h_2} \frac{\| e_p \|_2^2}{(w - \hat{w}_{|-1})^* \Pi_0^{-1} (w - \hat{w}_{|-1}) + \| v \|_2^2} \tag{22}$$

where Π_0 is a positive definite matrix that reflects apriori knowledge as to how close w is to the initial guess $\hat{w}_{|-1}$.

From root Fri Oct 29 16:55:23 1993 Received: from buckwheat.stanford.edu ([36.60.0.118]) by gibran (4.1/ECE.UCSB-v1.4) id AA00349; Fri, 29 Oct 93 16:55:22 PDT Received: from localhost (hassibi@localhost) by buckwheat.stanford.edu (8.6.1/8.6) id QAA16679 for sayed@gibran.ece.ucsb.edu; Fri, 29 Oct 1993 16:55:37 -0700 Date: Fri, 29 Oct 1993 16:55:37 -0700 From: Babak Hassibi ¡hassibi@buckwheat.stanford.edu¿ Message-Id: ¡199310292355.QAA16679@buckwheat.stanford.edu¿ To: sayed@gibran.ece.ucsb.edu Subject: adaline.tex for ima paper Content-Length: 1430 X-Lines: 30 Status: RO

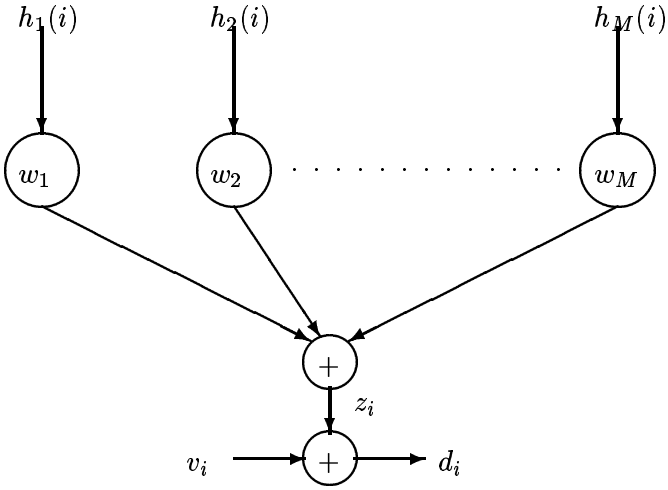

Figure 2: Signal model

¿From now on we shall assume, without loss of generality, that Π_0 has the special form $\Pi_0 = \mu I$, where μ is a positive constant.

Before closing this section we should remark that the conventional H^2 (or least squares) criterion recursively minimizes the following cost function:

$$\min_w \ (w - \hat{w}_{|-1})^* {\Pi_0}^{-1} (w - \hat{w}_{|-1}) + \sum_{j=0}^{i} |d(j) - h_j w|^2 \tag{23}$$

When w and the $\{v_j\}$ are independent Gaussian random variables with variances Π_0 and I respectively, the above criterion yields the maximum likelihood estimate of w. The recursive solution in its most most natural form, involves propagating a Riccati variable, yielding the so-called RLS algorithm. It has long been thought that LMS is an approximate algorithm where the Riccati variable is set equal to a constant matrix (most commonly, a multiple of the identity matrix), which leads to a simpler and faster algorithm.

However, we shall presntly see that the LMS algorithm, does in fact exactly minimize a *different* criterion, namely the H^∞ criterion.

Note that the H^∞ optimal adaptive filters guarantee the smallest estimation energy over all possible disturbances of fixed energy, and therefore will have better robust behaviour to disturbance variation. Moreover, in the special case when w and the $\{v_j\}$ are independent Gaussian random variables with variances Π_0 and I, respectively, we shall obtain an additional interpretation of the LMS algorithm, *viz.* it is an optimal risk-sensitive solution in the sense of Whittle.

5 Main Result

At this point we need one more definition.

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

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

5.1 The Normalized LMS Algorithm

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

Theorem 5 (Normalized LMS Algorithm) *Consider the state-space model (18), and suppose we want to minimize the H^∞ norm of the transfer operator $T_{f,i}$ from the unknowns w and $\{v_j\}_{j=0}^i$ to the filtered error $\{e_{f,j} = \hat{z}_{j|j} - h_j w\}_{j=0}^i$. If the input data $\{h_j\}$ is exciting, then the minimum H^∞ norm is*

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

In this case the central optimal H^∞ aposteriori 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:

$$\hat{w}_{|j+1} = \hat{w}_{|j} + \frac{\mu h_{j+1}^*}{1 + \mu h_{j+1} h_{j+1}^*}(d_{j+1} - h_{j+1}\hat{w}_{|j}) \quad , \quad \hat{w}_{|-1} \tag{24}$$

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 $\hat{w}_{|-1}$. 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

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

and the ratio in (21) can be made arbitrarily close to one.

The surprising fact though is that γ_{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.

Proof of Theorem 5: We shall use the aposteriori filter of Theorem 1 with $F_i = I$, $G_i = 0$, $H_i = h_i$, 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, 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

$$P_{i+1}^{-1} = \mu^{-1} I + (1-\gamma^{-2}) \sum_{j=0}^{i} h_j^* h_j \tag{25}$$

Now we need to check the existence condition (6) and find the optimum γ_{opt}. It follows from the above expression for P_{i+1}^{-1} that we have

$$P_{i+1}^{-1} + H_{i+1}^* H_{i+1} - \gamma^{-2} L_{i+1}^* L_{i+1} = \mu^{-1} I + (1-\gamma^{-2}) \sum_{j=0}^{i+1} h_j^* h_j \tag{26}$$

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

$$\sum_{j=0}^{i+1} |h_k(j)|^2 > \frac{\mu^{-1}}{\gamma^{-2} - 1}$$

This implies that the k^{th} diagonal entry of the matrix on the right hand side of (26) is negative, viz.,

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

Consequently, $P_{i+1}^{-1} + H_{i+1}^* H_{i+1} - \gamma^{-2} L_{i+1}^* L_{i+1}$ cannot be positive-definite. Therefore, $\gamma_{opt} \geq 1$. We now verify that γ_{opt} is indeed 1. For this purpose, we note that if we consider $\gamma = 1$ then from equation (25) we have $P_i = \mu I > 0$ for all i and the existence condition is satisfied. If we now write the aposteriori filter for $\gamma_{opt} = 1$, with $P_i = \mu I$, we get the desired so-called normalized LMS algorithm.

■

5.2 The LMS Algorithm

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

Theorem 6 (LMS Algorithm) *Consider the state-space model (18), and suppose we want to minimize the H^∞ norm of the transfer operator $T_{p,i}$ from the unknowns w and $\{v_j\}_{j=0}^i$ to the predicted error $\{e_{p,j} = \hat{z}_j - h_j w\}_{j=0}^i$. If the input data $\{h_j\}$ is exciting, and*

$$0 < \mu < \inf_i \frac{1}{h_i h_i^*} \tag{27}$$

then the minimum H^∞ norm is

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

In this case the central optimal apriori H^∞ filter is

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

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

$$\hat{w}_{|j} = \hat{w}_{|j-1} + \mu h_j^*(d_j - h_j \hat{w}_{|j-1}) \quad , \quad \hat{w}_{|-1} \tag{28}$$

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 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, (27) is satisfied, which leads to $\gamma_{opt} = 1$. Writing the H^∞ apriori filter equations for $\gamma = 1$ yields

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

■

The above result indicates that if the learning rate μ is chosen according to (27), 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 (27) with the bound studied in [2] and [21].

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 aposteriori and apriori 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.

5.3 Example

To illustrate the robustness of the LMS algorithm we consider a special case of model (18) where h_i is now a scalar that randomly takes on the values $+1$ and -1.

$$\begin{aligned} x_{i+1} &= x_i \quad , \quad x_0 = w \\ d_i &= h_i x_i + v_i \end{aligned} \tag{29}$$

Assuming we have observed N points of data, we can then use the LMS algorithm to write the transform operator $T_{lms,N}(\mu)$ that maps the disturbances $\{\mu^{-\frac{1}{2}}x_0, v_i\}$ to the $\{e_{p,i}\}$.

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

$$\underbrace{\begin{bmatrix} \mu^{\frac{1}{2}}h_0 & 0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}}(1-\mu)h_1 & -\mu h_1 h_0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}}(1-\mu)^2 h_2 & -\mu(1-\mu)h_2 h_0 & -\mu h_2 h_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu^{\frac{1}{2}}(1-\mu)^{N-1}h_{N-1} & -\mu(1-\mu)^{N-2}h_{N-1}h_0 & -\mu(1-\mu)^{N-3}h_{N-1}h_1 & \dots & -\mu h_{N-1}h_{N-2} \end{bmatrix}}_{T_{lms,N}(\mu)} \begin{bmatrix} \mu^{-\frac{1}{2}}x_0 \\ v_0 \\ \vdots \\ v_{N-2} \end{bmatrix} \tag{30}$$

Suppose now we use the RLS algorithm (*viz.* the Kalman filter) to estimate the states in (29), *i.e.*,

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

where

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

and

$$p_{i+1} = p_i - \frac{|h_i|^2 p_i^2}{1 + p_i|h_i|^2} = p_i - \frac{p_i^2}{1+p_i} = \frac{p_i}{1+p_i} \qquad , \qquad p_0 = \mu \tag{31}$$

then we can write the transfer operator $T_{rls,N}$ that maps the disturbances to the predicted errors as follows:

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

$$\underbrace{\begin{bmatrix} \mu^{\frac{1}{2}}h_0 & 0 & 0 & \dots & 0 \\ \mu^{\frac{1}{2}}\frac{h_1}{1+\mu} & -\mu\frac{h_1 h_0}{1+\mu} & 0 & \dots & 0 \\ \mu^{\frac{1}{2}}\frac{h_2}{1+2\mu} & -\mu\frac{h_2 h_0}{1+2\mu} & -\mu\frac{h_2 h_1}{1+2\mu} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu^{\frac{1}{2}}\frac{h_{N-1}}{1+(N-1)\mu} & -\mu\frac{h_{N-1}h_0}{1+(N-1)\mu} & -\mu\frac{h_{N-1}h_1}{1+(N-1)\mu} & \dots & -\mu\frac{h_{N-1}h_{N-2}}{1+(N-1)\mu} \end{bmatrix}}_{T_{rls,N}(\mu)} \begin{bmatrix} \mu^{-\frac{1}{2}}x_0 \\ v_0 \\ \vdots \\ v_{N-2} \end{bmatrix} \tag{32}$$

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 (27) 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 3 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.

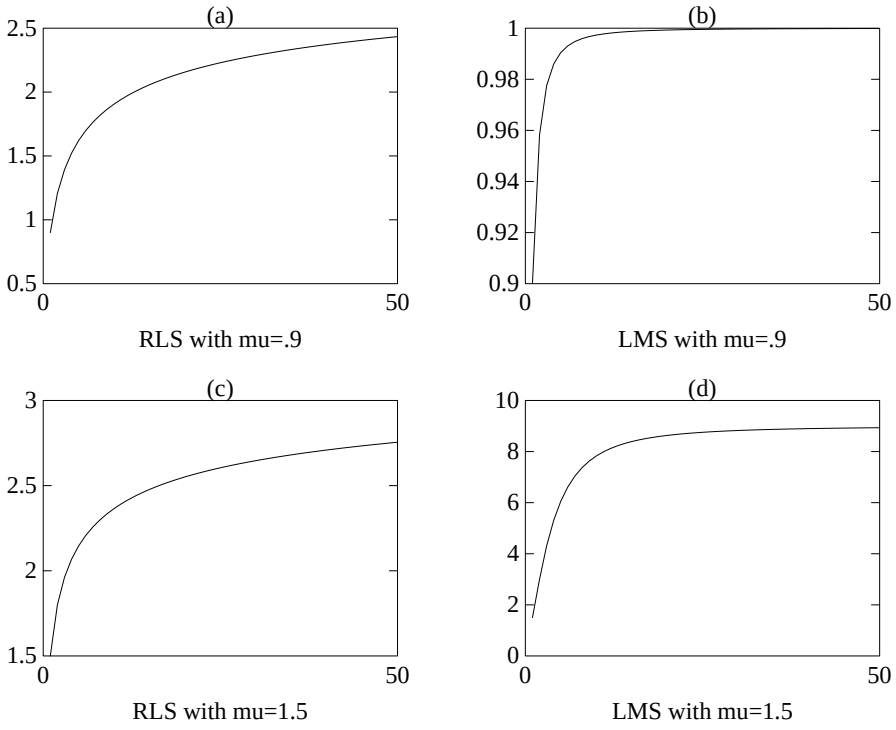

Figure 3: Maximum singular value of transfer operators $T_{lms,N}(\mu)$ and $T_{rls,N}(\mu)$ as a function of N for the values $\mu = .9$ and $\mu = 1.5$.

Figure 4 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 4a. As can be seen from Figure 4b 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 4c, 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.

5.4 Discussion

In Section 5.1 we motivated the $\gamma_{opt} = 1$ 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 aposteriori 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.*,

$$\hat{z}_{i|i} = d_i \tag{33}$$

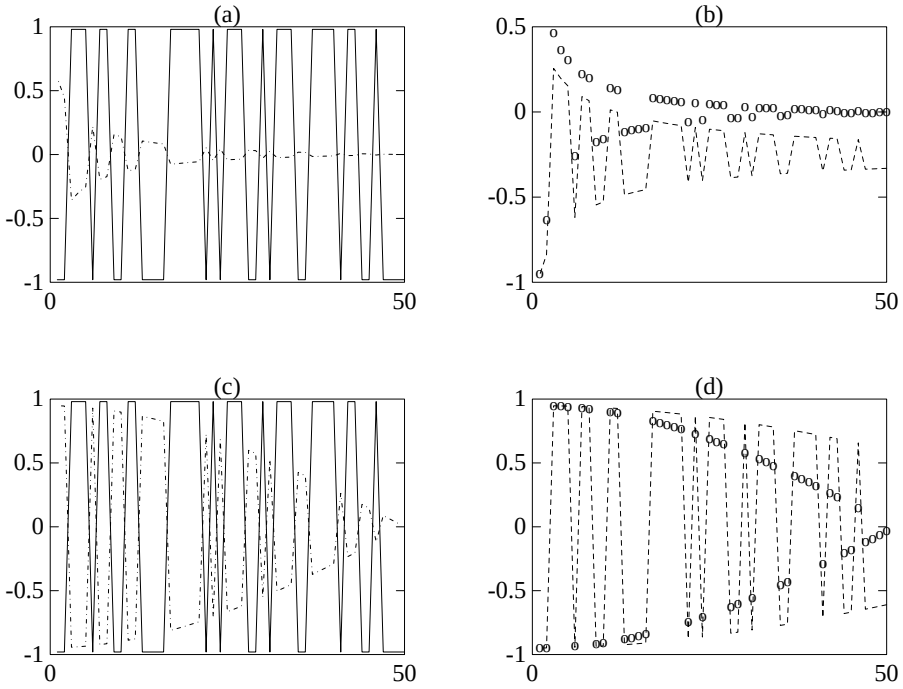

Figure 4: Worst case disturbances and the corresponding predicted errors for RLS and LMS. (a) The solid line represents the uncorrupted output $h_i x_i$ and the dashed line represents the worst case RLS disturbance. (b) The dashed line and the dotted line represent the RLS and LMS predicted errors, respectively, 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.

The corresponding filtered error is:

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

Thus the ratio in (21) can be made arbitrarily close to one, and the estimator (33) will achieve the same $\gamma_{opt} = 1$ that the normalized LMS algorithm does.

Formally, the estimator (33) may be obtained from the normalized LMS algorithm (24) by letting $\mu \to \infty$. However, (33) will achieve $\gamma_{opt} = 1$ for any value of μ.

The fact that the simplistic estimator (33) (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 sensitivity of the RLS algorithm to different disturbance signals, as illustrated in the example of Section 5.3, 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 6 we shall develop the full parametrization of all H^∞ optimal aposteriori and apriori adaptive filters, and show how to obtain (33) as a special case of this parametrization. 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 shall discuss in Section 7: they are risk-sensitive optimal (i.e. they optimize a certain exponential cost criterion) and can also be shown to be maximum entropy.

The main problem with the estimator (33) is that it makes absolutely no use of the state-space model (18). We should note that it is not possible to come up with such a simple minded estimator in the apriori case: indeed as we shall see in the next section, the apriori estimator corresponding to (33) is highly nontrivial. The reason seems to be that since in the apriori case one deals with predicted error energy, it is inevitable that one must make use of the state-space model (18) in order to construct an optimal prediction of the *next* output. Thus in the apriori case, the problems arising from such unreasonable estimators such as (33) are avoided.

6 All H^∞ Adaptive Filters

In Section 5.4 we came up with an alternative optimal H^∞ aposteriori filter. We shall presently use the results of Theorems 3 and 4 to parametrize all optimal H^∞ apriori and aposteriori filters.

Theorem 7 (All H^∞ Aposteriori Adaptive Filters) *All H^∞ optimal aposteriori adaptive filters that achieve $\gamma_{opt} = 1$ are given by*

$$\begin{aligned} \hat{z}_{i|i} &= h_i \hat{x}_{i|i} + (I + \mu h_i h_i^*)^{-\frac{1}{2}} \\ &\quad \mathcal{S}_i \left((I + \mu h_i h_i^*)^{\frac{1}{2}} (d_i - h_i \hat{x}_{i|i}), \ldots, (I + \mu h_0 h_0^*)^{\frac{1}{2}} (d_0 - h_0 \hat{x}_{0|0}) \right) \end{aligned} \tag{34}$$

where $\hat{x}_{i|i}$ is the estimated state of the normalized LMS algorithm with parameter μ, and

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

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

$$\sum_{j=0}^{i} |\mathcal{S}_j(a_j, \ldots, a_0)|^2 \leq \sum_{j=0}^{i} |a_j|^2$$

Proof: Using the result of Theorem 3 with $H_i = h_i$ and $L_i = h_i$, the full parametrization of all H^∞ aposteriori adaptive filters is given by

$$\begin{aligned} \hat{z}_{i|i} &= h_i \hat{x}_{i|i} + [I - h_i(P_i^{-1} + h_i^* h_i)^{-1} h_i^*]^{\frac{1}{2}} \\ &\quad \mathcal{S}_i \left((I + h_i P_i h_i^*)^{\frac{1}{2}} (d_i - h_i \hat{x}_{i|i}), \ldots, (I + h_0 P_0 h_0^*)^{\frac{1}{2}} (d_0 - h_0 \hat{x}_{0|0}) \right) \end{aligned} \tag{35}$$

Now from the proof of Theorem 5 we know that for all aposteriori filters that achieve $\gamma_{opt} = 1$, we have $P_i = \mu I$. Moreover we have the identity

$$I - h_i(P_i^{-1} + h_i^* h_i)^{-1} h_i^* = (I + h_i P_i h_i^*)^{-1}$$

Replacing the above expression along with $P_i = \mu I$ into (35) yields the desired result. ■

At this point we should note the significance of some special choices for the causal contraction $\mathcal{S}$.

- $\mathcal{S} = 0$: This yields the normalized LMS algorithm.
- $\mathcal{S} = I$: This yields

  $$\hat{z}_{i|i} = h_i \hat{x}_{i|i} + (I + \mu h_i h_i^*)^{-\frac{1}{2}} (I + \mu h_i h_i^*)^{\frac{1}{2}} (d_i - h_i \hat{x}_{i|i}) = d_i$$

  which is the simple minded estimator of Section 5.4.
- $\mathcal{S} = -I$: This yields

  $$\hat{z}_{i|i} = h_i \hat{x}_{i|i} - (I + \mu h_i h_i^*)^{-\frac{1}{2}} (I + \mu h_i h_i^*)^{\frac{1}{2}} (d_i - h_i \hat{x}_{i|i}) = 2h_i \hat{x}_{i|i} - d_i$$

Thus it is quite obvious that the different H^∞ optimal aposteriori adaptive filters may have quite different behaviour with respect to other desirable criteria.

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

$$\begin{aligned} \hat{z}_i &= h_i \hat{x}_i + (I - \mu h_i h_i^*)^{\frac{1}{2}} \\ &\quad \mathcal{S}_i \left((I - \mu h_{i-1} h_{i-1}^*)^{\frac{1}{2}} (d_{i-1} - h_{i-1} \hat{x}_{i-1}), \ldots, (I - \mu h_0 h_0^*)^{\frac{1}{2}} (d_0 - h_0 \hat{x}_0) \right) \end{aligned} \tag{36}$$

where $\hat{x}_i$ is the state estimate of the LMS algorithm with learning rate μ, and $\mathcal{S}$ is any (possibly nonlinear) contractive causal mapping.

Proof: Using the result of Theorem 4 with $H_i = h_i$ and $L_i = h_i$, the full parametrization of all H^∞ apriori adaptive filters is given by

$$\begin{aligned} \hat{z}_i &= h_i\hat{x}_i + [I - h_iP_ih_i^*]^{\frac{1}{2}} \\ &\quad \mathcal{S}_i\left((I + h_{i-1}\tilde{P}_{i-1}h_{i-1}^*)^{-\frac{1}{2}}(d_{i-1} - h_{i-1}\hat{x}_{i-1}), \ldots, (I + h_0\tilde{P}_0h_0^*)^{-\frac{1}{2}}(d_0 - h_0\hat{x}_0)\right) \end{aligned} \tag{37}$$

where $\tilde{P}_i = (P_i^{-1} - h_i^*h_i)^{-1}$. Now from the proof of Theorem 6 we know that for all apriori filters that achieve $\gamma_{opt} = 1$, we have $P_i = \mu I$. Moreover we have the identity

$$I + h_i\tilde{P}_ih_i^* = I + h_i(P_i^{-1} - h_i^*h_i)^{-1}h_i = (I - h_iP_ih_i^*)^{-1}$$

Replacing the above expression along with $P_i = \mu I$ into (37) yields the desired result. ■

It is once more interesting to note the consequences of some special choices of the causal contraction $\mathcal{S}$.

- $\mathcal{S} = 0$: This yields the LMS algorithm.
- $\mathcal{S} = I$: This yields

  $$\hat{z}_i = h_i\hat{x}_i + (I - \mu h_ih_i^*)^{\frac{1}{2}}(I - \mu h_{i-1}h_{i-1}^*)^{\frac{1}{2}}(d_{i-1} - h_{i-1}\hat{x}_{i-1})$$

  which is the apriori adaptive filter that corresponds to the simple minded estimator of Section 5.4. Note that in this case the filter is highly nontrivial.
- $\mathcal{S} = -I$: This yields

  $$\hat{z}_i = h_i\hat{x}_i - (I - \mu h_ih_i^*)^{\frac{1}{2}}(I - \mu h_{i-1}h_{i-1}^*)^{\frac{1}{2}}(d_{i-1} - h_{i-1}\hat{x}_i)$$

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

7 Risk-Sensitive Optimality

In this section we shall 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 the recent textbook [16].

7.1 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 2, we shall 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:

$$\begin{aligned} x_{i+1} &= x_i \quad , \qquad x_0 = w \\ d_i &= h_ix_i + v_i \end{aligned} \tag{38}$$

where w and the $\{v_i\}$ are now independent Gaussian random variables with covariances Π_0 and I, respectively. Moreover, w is assumed to have mean $\hat{w}_{|-1}$, and the $\{v_i\}$ are assumed to be zero mean. 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 $\hat{z}_i = \mathcal{F}_f(d_0, d_1, \ldots, d_{i-1})$ 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.* [22, 23, 20]):

$$\min_{\{\hat{z}_j\}} E \sum_{j=0}^{i} e_{p,j}^* e_{p,j} = \min_{\{\hat{z}_j\}} E \parallel e_p \parallel^2 \tag{39}$$

where the expectation is taken over the Gaussian random variables w and $\{v_j\}$ 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 (39) is known as the *risk-neutral* filter.

An alternative criterion that is risk-sensitive has been extensively studied in [14, 15, 16, 17] and corresponds to the following minimization problem

$$\min_{\{\hat{z}_{j|j}\}} \mu_{f,i}(\theta) = \min_{\{\hat{z}_{j|j}\}} \left(-\frac{2}{\theta} log\left[E exp(-\frac{\theta}{2}\mathbf{C}_{f,i})\right]\right) \tag{40a}$$

or

$$\min_{\{\hat{z}_j\}} \mu_{p,i}(\theta) = \min_{\{\hat{z}_j\}} \left(-\frac{2}{\theta} log\left[E exp(-\frac{\theta}{2}\mathbf{C}_{p,i})\right]\right) \tag{40b}$$

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 (40a) and (40b) are known as the aposteriori and apriori *exponential cost functions*, and any filters that minimize $\mu_{f,i}(\theta)$ and $\mu_{p,i}(\theta)$ are referred to as aposteriori and apriori 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 $E exp(-\frac{\theta}{2}\mathbf{C}_i)$, 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 $E exp(-\frac{\theta}{2}\mathbf{C}_i)$, 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 [24] and has been further discussed in [16, 13]. 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 $\gamma_{opt} = -\bar{\theta}^{-\frac{1}{2}}$.

7.2 Risk-sensitive Adaptive Filtering

Using the discussion of Section 7.1, 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 (38) 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*

$$\min_{\{\hat{z}_{j|j}\}} \mu_{f,i}(\theta) = \min_{\{\hat{z}_{j|j}\}} \left(2log \left[Eexp(\frac{1}{2}\mathbf{C}_{f,i}) \right] \right) \tag{41}$$

where $\mathbf{C}_{f,i} = \sum_{j=0}^{i} e_{f,i}^* e_{f,i}$, 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

$$\hat{w}_{|i+1} = \hat{w}_{|i} + \frac{\mu h_{i+1}^*}{1 + \mu h_{i+1} h_{i+1}^*}(d_{i+1} - h_{i+1}\hat{w}_{|i}) \quad , \quad \hat{w}_{|-1} \tag{42}$$

Theorem 10 (LMS and Risk-sensitivity) *Consider the state-space model (38) 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

$$\min_{\{\hat{z}_j\}} \mu_{p,i}(\theta) = \min_{\{\hat{z}_j\}} \left(2log \left[Eexp(\frac{1}{2}\mathbf{C}_{p,i}) \right] \right) \tag{43}$$

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

$$\hat{w}_{|i} = \hat{w}_{|i-1} + \mu h_i^*(d_i - h_i \hat{w}_{|i-1}) \quad , \quad \hat{w}_{|-1} \tag{44}$$

Thus LMS and normalized LMS are risk-averse filters that avoid the occasional occurrence of large estimation error energies at the price of tolerating the frequent occurrence of moderate ones. This fact is in agreement with the intuition we have gained from the H^∞ optimality of these algorithms.

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. Morever, among all H^∞ estimators that achieve a certain level γ, the central solution can be shown to be the maximum entropy [18] solution. However, we shall not pursue these directions here.

8 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 Prof. L. Ljung for contributing to the discussion in Section 5.4.

References

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

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] I. Yaesh and U. Shaked. $H^\infty-$optimal estimation: The discrete time case. In *Proc. Inter. Symp. on MTNS*, pages 261–267, Kobe, Japan, June 1991.

[11] 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.

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

[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] 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] 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.

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

[20] A. H. Sayed and T. Kailath. A state-space approach to adaptive filtering. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, MN, 1993.

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

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

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

[24] 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.