Sayed, A.H. & Rupp, M. “Robustness Issues in Adaptive Filtering”
Digital Signal Processing Handbook
Ed. Vijay K. Madisetti and Douglas B. Williams
Boca Raton: CRC Press LLC, 1999

20

Robustness Issues in Adaptive Filtering

Ali H. Sayed
University of California, Los Angeles

Markus Rupp
Bell Laboratories
Lucent Technologies

Adaptive filters are systems that adjust themselves to a changing environment. They are designed to meet certain performance specifications and are expected to perform reasonably well under the operating conditions for which they have been designed. In practice, however, factors that may have been ignored or overlooked in the design phase of the system can affect the performance of the adaptive scheme that has been chosen for the system. Such factors include unmodeled dynamics, modeling errors, measurement noise, and quantization errors, among others, and their effect on the performance of an adaptive filter could be critical to the proposed application. Moreover, technological advancements in digital circuit and VLSI design have spurred an increase in the range of new adaptive filtering applications in fields ranging from biomedical engineering to wireless communications. For these new areas, it is increasingly important to design adaptive schemes that are tolerant to unknown or nontraditional factors and effects. The aim of this chapter is to explore and determine the robustness properties of some classical adaptive schemes. Our presentation is meant as an introduction to these issues, and many of the relevant details of specific topics discussed in this section, and alternative points of view, can be found in the references at the end of the chapter.

20.1 Motivation and Example

A classical application of adaptive filtering is that of system identification. The basic problem formulation is depicted in Fig. 20.1, where z^{-1} denotes the unit-time delay operator. The diagram contains two system blocks: one representing the *unknown plant* or system and the other containing

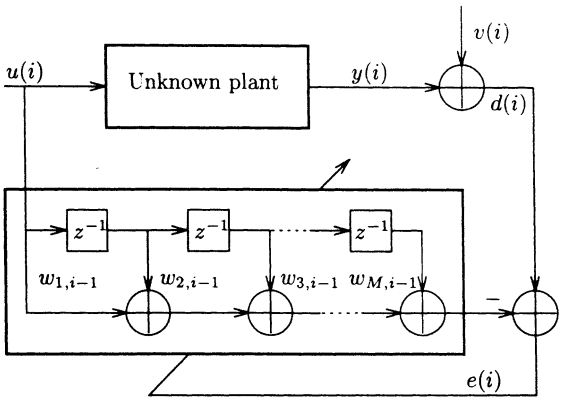


FIGURE 20.1: A system identification example.

a time-variant tapped-delay-line or finite-impulse-response (FIR) filter structure. The unknown plant represents an arbitrary relationship between its input and output. This block might implement a pole-zero transfer function, an all-pole or *autoregressive* transfer function, a fixed or time-varying FIR system, a nonlinear mapping, or some other complex system. In any case, it is desired to determine an FIR model for the unknown system of a predetermined impulse response length M, and whose coefficients at time $i-1$ are denoted by $\{w_{1,i-1}, w_{2,i-1}, \ldots, w_{M,i-1}\}$. The unknown system and the FIR filter are excited by the same input sequence $\{u(i)\}$, where the time origin is at $i = 0$.

If we collect the FIR coefficients into a column vector, say $\mathbf{w}_{i-1} = \text{col}\{w_{1,i-1}, w_{2,i-1}, \ldots, w_{M,i-1}\}$, and define the state vector of the FIR model at time i as $\mathbf{u}_i = \text{col}\{u(i), u(i-1), \ldots, u(i-M+1)\}$, then the output of the FIR filter at time i is the inner product $\mathbf{u}_i^T \mathbf{w}_{i-1}$. In principle, this inner product should be compared with the output $y(i)$ of the unknown plant in order to determine whether or not the FIR output is a good enough approximation for the output of the plant and, therefore, whether or not the current coefficient vector $\mathbf{w}_{i-1}$ should be updated.

In general, however, we do not have direct access to the uncorrupted output $y(i)$ of the plant but rather to a noisy measurement of it, say $d(i) = y(i) + v(i)$. The purpose of an adaptive scheme is to employ the output error sequence $\{e(i) = d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1}\}$, which measures how far $d(i)$ is from $\mathbf{u}_i^T \mathbf{w}_{i-1}$, in order to update the entries of $\mathbf{w}_{i-1}$ and provide a better model, say $\mathbf{w}_i$, for the unknown system. That is, the purpose of the adaptive filter is to employ the available data at time i, $\{d(i), \mathbf{w}_{i-1}, \mathbf{u}_i\}$, in order to update the coefficient vector $\mathbf{w}_{i-1}$ into a presumably better estimate vector $\mathbf{w}_i$.

In this sense, we may regard the adaptive filter as a recursive estimator that tries to come up with a coefficient vector $\mathbf{w}$ that "best" matches the observed data $\{d(i)\}$ in the sense that, for all i, $d(i) \approx \mathbf{u}_i^T \mathbf{w} + v(i)$ to good accuracy. The successive $\mathbf{w}_i$ provide estimates for the unknown and desired $\mathbf{w}$.

20.2 Adaptive Filter Structure

We may reformulate the above adaptive problem in mathematical terms as follows. Let $\{\mathbf{u}_i\}$ be a sequence of *regression vectors* and let $\mathbf{w}$ be an unknown column vector to be estimated or identified. Given noisy measurements $\{d(i)\}$ that are assumed to be related to $\mathbf{u}_i^T \mathbf{w}$ via an additive noise model of the form

$$d(i) = \mathbf{u}_i^T \mathbf{w} + v(i) \ , \tag{20.1}$$

we wish to employ the given data $\{d(i), \mathbf{u}_i\}$ in order to provide recursive estimates for $\mathbf{w}$ at successive time instants, say $\{\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots\}$. We refer to these estimates as *weight* estimates since they provide estimates for the coefficients or weights of the tapped-delay model.

Most adaptive schemes perform this task in a recursive manner that fits into the following general description: starting with an initial guess for $\mathbf{w}$, say $\mathbf{w}_{-1}$, iterate according to the learning rule

$$\begin{pmatrix} \text{new weight} \\ \text{estimate} \end{pmatrix} = \begin{pmatrix} \text{old weight} \\ \text{estimate} \end{pmatrix} + \begin{pmatrix} \text{correction} \\ \text{term} \end{pmatrix},$$

where the correction term is usually a function of $\{d(i), \mathbf{u}_i, \text{old weight estimate}\}$. More compactly, we may write $\mathbf{w}_i = \mathbf{w}_{i-1} + f[d(i), \mathbf{u}_i, \mathbf{w}_{i-1}]$, where $\mathbf{w}_i$ denotes an estimate for $\mathbf{w}$ at time i and f denotes a function of the data $\{d(i), \mathbf{u}_i, \mathbf{w}_{i-1}\}$ or of previous values of the data, as in the case where only a filtered version of the error signal $d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1}$ is available. In this context, the well-known least-mean-square (LMS) algorithm has the form

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu \cdot \mathbf{u}_i \cdot [d(i) - \mathbf{u}_i^T \cdot \mathbf{w}_{i-1}], \tag{20.2}$$

where μ is known as the *step-size* parameter.

20.3 Performance and Robustness Issues

The performance of an adaptive scheme can be studied from many different points of view. One distinctive methodology that has attracted considerable attention in the adaptive filtering literature is based on stochastic considerations that have become known as the *independence assumptions.* In this context, certain statistical assumptions are made on the natures of the noise signal $\{v(i)\}$ and of the regression vectors $\{\mathbf{u}_i\}$, and conclusions are derived regarding the steady-state behavior of the adaptive filter.

The discussion in this chapter avoids statistical considerations and develops the analysis in a purely deterministic framework that is convenient when prior statistical information is unavailable or when the independence assumptions are unreasonable. The conclusions discussed herein highlight certain features of the adaptive algorithms that hold regardless of any statistical considerations in an adaptive filtering task.

Returning to the data model in (20.1), we see that it assumes the existence of an unknown weight vector $\mathbf{w}$ that describes, along with the regression vectors $\{\mathbf{u}_i\}$, the uncorrupted data $\{y(i)\}$. This assumption may or may not hold.

For example, if the unknown plant in the system identification scenario of Fig. 20.1 is itself an FIR system of length M, then there exists an unknown weight vector $\mathbf{w}$ that satisfies (20.1). In this case, the successive estimates provided by the adaptive filter attempt to identify the unknown weight vector of the plant.

If, on the other hand, the unknown plant of Fig. 20.1 is an autoregressive model of the simple form

$$\frac{1}{1 - cz^{-1}} = 1 + cz^{-1} + c^2 z^{-2} + c^3 z^{-3} + \ldots$$

where $|c| < 1$, then an infinitely long tapped-delay line is necessary to justify a model of the form (20.1). In this case, the first term in the linear regression model (20.1) for a finite order M cannot describe the uncorrupted data $\{y(i)\}$ exactly, and thus modeling errors are inevitable. Such modeling errors can naturally be included in the noise term $v(i)$. Thus, we shall use the term $v(i)$ in (20.1) to account not only for measurement noise but also for modeling errors, unmodeled dynamics, quantization effects, and other kind of disturbances within the system. In many cases,

the performance of the adaptive filter depends on how these unknown disturbances affect the weight estimates.

A second source of error in the adaptive system is due to the initial guess $\mathbf{w}_{-1}$ for the weight vector. Due to the iterative nature of our chosen adaptive scheme, it is expected that this initial weight vector plays less of a role in the steady-state performance of the adaptive filter. However, for a finite number of iterations of the adaptive algorithm, both the noise term $v(i)$ and the initial weight error vector $(\mathbf{w} - \mathbf{w}_{-1})$ are disturbances that affect the performance of the adaptive scheme, particularly since the system designer often has little control over them.

The purpose of a robust adaptive filter design, then, is to develop a recursive estimator that minimizes in some well-defined sense the effect of any unknown disturbances on the performance of the filter. For this purpose, we first need to quantify or measure the effect of the disturbances. We address this concern in the following sections.

20.4 Error and Energy Measures

Assuming that the model (20.1) is reasonable, two error quantities come to mind. The first one measures how far the weight estimate $\mathbf{w}_{i-1}$ provided by the adaptive filter is from the true weight vector $\mathbf{w}$ that we are trying to identify. We refer to this quantity as the weight error at time $(i-1)$, and we denote it by $\tilde{\mathbf{w}}_{i-1} = \mathbf{w} - \mathbf{w}_{i-1}$. The second type of error measures how far the estimate $\mathbf{u}_i^T \mathbf{w}_{i-1}$ is from the uncorrupted output term $\mathbf{u}_i^T \mathbf{w}$. We shall call this the *a priori estimation error*, and we denote it by $e_a(i) = \mathbf{u}_i^T \tilde{\mathbf{w}}_{i-1}$. Similarly, we define an *a posteriori estimation error* as $e_p(i) = \mathbf{u}_i^T \tilde{\mathbf{w}}_i$. Comparing with the definition of the *a priori* error, the *a posteriori* error employs the most recent weight error vector.

Ideally, one would like to make the estimation errors $\{\tilde{\mathbf{w}}_i, e_a(i)\}$ or $\{\tilde{\mathbf{w}}_i, e_p(i)\}$ as small as possible. This objective is hindered by the presence of the disturbances $\{\tilde{\mathbf{w}}_{-1}, v(i)\}$. For this reason, an adaptive filter is said to be *robust* if the effects of the disturbances $\{\tilde{\mathbf{w}}_{-1}, v(i)\}$ on the resulting estimation errors $\{\tilde{\mathbf{w}}_i, e_a(i)\}$ or $\{\tilde{\mathbf{w}}_i, e_p(i)\}$ is small in a well-defined sense. To this end, we can employ one of several measures to denote how "small" these effects are. For our discussion, a quantity known as the *energy* of a signal will be used to quantify these effects. The energy of a sequence $x(i)$ of length N is measured by $\mathcal{E}_x = \sum_{i=0}^{N-1} |x(i)|^2$. A finite energy sequence is one for which $\mathcal{E}_x < \infty$ as $N \to \infty$. Likewise, a finite power sequence is one for which

$$\mathcal{P}_x = \lim_{N\to\infty} \left(\frac{1}{N} \sum_{i=0}^{N-1} |x(i)|^2 \right) < \infty .$$

20.5 Robust Adaptive Filtering

We can now quantify what we mean by robustness in the adaptive filtering context. Let $\mathcal{A}$ denote any adaptive filter that operates causally on the input data $\{d(i), \mathbf{u}_i\}$. A causal adaptive scheme produces a weight vector estimate at time i that depends only on the data available up to and including time i. This adaptive scheme receives as input the data $\{d(i), \mathbf{u}_i\}$ and provides as output the weight vector estimates $\{\mathbf{w}_i\}$. Based on these estimates, we introduce one or more estimation error quantities such as the pair $\{\tilde{\mathbf{w}}_{i-1}, e_a(i)\}$ defined above. Even though these quantities are not explicitly available because $\mathbf{w}$ is unknown, they are of interest to us as their magnitudes determine how well or how poorly a candidate adaptive filtering scheme might perform.

Figure 20.2 indicates the relationship between $\{d(i), \mathbf{u}_i\}$ to $\{\tilde{\mathbf{w}}_{i-1}, e_a(i)\}$ in block diagram form. This schematic representation indicates that an adaptive filter $\mathcal{A}$ operates on $\{d(i), \mathbf{u}_i\}$ and that

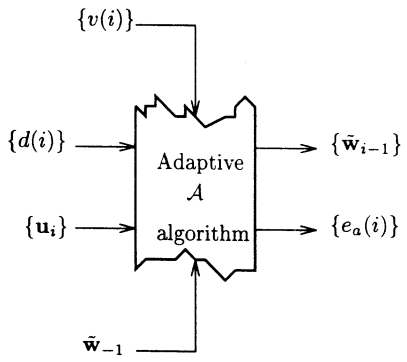


FIGURE 20.2: Input-output map of a generic adaptive scheme.

its performance relies on the sizes of the error quantities $\{\tilde{\mathbf{w}}_{i-1}, e_a(i)\}$, which could be replaced by the error quantities $\{\tilde{\mathbf{w}}_i, e_p(i)\}$ if desired. This representation explicitly denotes the quantities $\{\tilde{\mathbf{w}}_{-1}, v(i)\}$ as disturbances to the adaptive scheme.

In order to measure the effect of the disturbances on the performance of an adaptive scheme, it will be helpful to determine the explicit relationship between the disturbances and the estimation errors that is provided by the adaptive filter. For example, we would like to know what effect the noise terms and the initial weight error guess $\{\tilde{\mathbf{w}}_{-1}, v(i)\}$ would have on the resulting *a priori* estimation errors and the final weight error, $\{e_a(i), \tilde{\mathbf{w}}_N\}$, for a given adaptive scheme. Knowing such a relationship, we can then quantify the robustness of the adaptive scheme by determining the degree to which disturbances affect the size of the estimation errors.

We now illustrate how this disturbances-to-estimation-errors relationship can be determined by considering the LMS algorithm in (20.2). Since $d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1} = e_a(i) + v(i)$, we can subtract $\mathbf{w}$ from both sides of (20.2) to obtain the weight-error update equation

$$\tilde{\mathbf{w}}_i = \tilde{\mathbf{w}}_{i-1} - \mu \cdot \mathbf{u}_i \cdot [e_a(i) + v(i)] \,. \tag{20.3}$$

Assume that we run N steps of the LMS recursion starting with an initial guess $\tilde{\mathbf{w}}_{-1}$. This operation generates the weight error estimates $\{\tilde{\mathbf{w}}_0, \tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_N\}$ and the *a priori* estimation errors $\{e_a(0), \ldots, e_a(N)\}$.

Define the following two column vectors:

$$\underline{\text{dist}} = \text{col}\left\{\frac{1}{\sqrt{\mu}}\tilde{\mathbf{w}}_{-1}, v(0), v(1), \ldots, v(N)\right\} \;, \quad \underline{\text{error}} = \text{col}\left\{e_a(0), e_a(1), \ldots, e_a(N), \frac{1}{\sqrt{\mu}}\tilde{\mathbf{w}}_N\right\} \;.$$

The vector $\underline{\text{dist}}$ contains the disturbances that affect the performance of the adaptive filter. The initial weight error vector is scaled by $\mu^{-1/2}$ for convenience. Likewise, the vector $\underline{\text{error}}$ contains the *a priori* estimation errors and the final weight error vector which has also been scaled by $\mu^{-1/2}$. The weight error update relation in (20.3) allows us to relate the entries of both vectors in a straightforward manner. For example,

$$e_a(0) = \mathbf{u}_0^T \tilde{\mathbf{w}}_{-1} = \left(\sqrt{\mu}\,\mathbf{u}_0^T\right)\left(\frac{1}{\sqrt{\mu}}\,\tilde{\mathbf{w}}_{-1}\right) \;,$$

which shows how the first entry of $\underline{\text{error}}$ relates to the first entry of $\underline{\text{dist}}$. Similarly, for $e_a(1) = \mathbf{u}_1^T \tilde{\mathbf{w}}_0$ we obtain

$$e_a(1) = \left(\sqrt{\mu}\mathbf{u}_1^T [I - \mu \mathbf{u}_0 \mathbf{u}_0^T]\right) \frac{1}{\sqrt{\mu}}\tilde{\mathbf{w}}_{-1} \;-\; \left(\mu \mathbf{u}_1^T \mathbf{u}_0\right) v(0) \;,$$

which relates $e_a(1)$ to the first two entries of the vector $\underline{\text{dist}}$. Continuing in this manner, we can relate $e_a(2)$ to the first three entries of $\underline{\text{dist}}$, $e_a(3)$ to the first four entries of $\underline{\text{dist}}$, and so on.

In general, we can compactly express this relationship as

$$\underbrace{\begin{bmatrix} e_a(0) \\ e_a(1) \\ \vdots \\ \hline e_a(N) \\ \hline \frac{1}{\sqrt{\mu}}\tilde{\mathbf{w}}_N \end{bmatrix}}_{\underline{\text{error}}} = \underbrace{\begin{bmatrix} \times & & & & & \\ \times & \times & & & O & \\ \vdots & & & \ddots & & \\ \times & \times & \times & \times & \times & \times \end{bmatrix}}_{\mathcal{T}} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{\mu}}\tilde{\mathbf{w}}_{-1} \\ \hline v(0) \\ v(1) \\ \vdots \\ v(N) \end{bmatrix}}_{\underline{\text{dist}}}$$

where the symbol $\times$ is used to denote the entries of the lower triangular mapping $\mathcal{T}$ relating $\underline{\text{dist}}$ to $\underline{\text{error}}$. The specific values of the entries of $\mathcal{T}$ are not of interest for now, although we have indicated how the expressions for these $\times$ terms can be found. However, the causal nature of the adaptive algorithm requires that $\mathcal{T}$ be of lower triangular form.

Given the above relationship, our objective is to quantify the effect of the disturbances on the estimation errors. Let $\mathcal{E}_d$ and $\mathcal{E}_e$ denote the energies of the vectors $\underline{\text{dist}}$ and $\underline{\text{error}}$, respectively, such that

$$\mathcal{E}_e = \frac{1}{\mu}\|\tilde{\mathbf{w}}_N\|^2 + \sum_{i=0}^{N}|e_a(i)|^2 \quad \text{and} \quad \mathcal{E}_d = \frac{1}{\mu}\|\tilde{\mathbf{w}}_{-1}\|^2 + \sum_{i=0}^{N}|v(i)|^2 ,$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. We shall say that the LMS adaptive algorithm is *robust with level* γ if a relation of the form

$$\frac{\mathcal{E}_e}{\mathcal{E}_d} \leq \gamma^2 , \tag{20.4}$$

holds for some positive γ and for *any* nonzero, finite-energy disturbance vector $\underline{\text{dist}}$. In other words, no matter what the disturbances $\{\tilde{\mathbf{w}}_{-1}, v(i)\}$ are, the energy of the resulting estimation errors will never exceed γ^2 times the energy of the associated disturbances.

The form of the mapping $\mathcal{T}$ affects the value of γ in (20.4) for any particular algorithm. To see this result, recall that for any finite-dimensional matrix A, its maximum singular value, denoted by $\bar{\sigma}(A)$, is defined by $\bar{\sigma}(A) = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$. Hence, the square of the maximum singular value, $\bar{\sigma}^2(A)$, measures the maximum energy gain from the vector x to the resulting vector Ax. Therefore, if a relation of the form (20.4) should hold for any nonzero disturbance vector $\underline{\text{dist}}$, then it means that

$$\max_{\underline{\text{dist}}\neq 0} \frac{\|\,\mathcal{T}\underline{\text{dist}}\,\|}{\|\,\underline{\text{dist}}\,\|} \leq \gamma .$$

Consequently, the maximum singular value of $\mathcal{T}$ must be bounded by γ. This imposes a condition on the allowable values for γ; its smallest value cannot be smaller than the maximum singular value of the resulting $\mathcal{T}$.

Ideally, we would like the value of γ in (20.4) to be as small as possible. In particular, an algorithm for which the value of γ is 1 would guarantee that the estimation error energy will never exceed the disturbance energy, no matter what the natures of the disturbances are! Such an algorithm would possess a good degree of robustness since it would guarantee that the disturbance energy will never be unnecessarily magnified.

Before continuing our study, we ask and answer the obvious questions that arise at this point:

- *What is the smallest possible value for γ for the LMS algorithm*? It turns out for the LMS algorithm that, under certain conditions on the step-size parameter, the smallest possible value for γ is 1. Thus, $\mathcal{E}_e \leq \mathcal{E}_d$ for the LMS algorithm.
- *Does there exist any other causal adaptive algorithm that would result in a value for γ in (20.4) that is smaller than one*? It can be argued that no such algorithm exists for the model (20.1) and criterion (20.4).

In other words, the LMS algorithm is in fact the most robust adaptive algorithm in the sense defined by (20.4). This result provides a rigorous basis for the excellent robustness properties that the LMS algorithm, and several of its variants, have shown in practical situations. The references at the end of the chapter provide an overview of the published works that have established these conclusions. Here, we only motivate them from first principles. In so doing, we shall also discuss other results (and tools) that can be used in order to impose certain robustness and convergence properties on other classes of adaptive schemes.

20.6 Energy Bounds and Passivity Relations

Consider the LMS recursion in (20.2), with a time-varying step-size $\mu(i)$ for purposes of generality, as given by

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu(i) \cdot \mathbf{u}_i \cdot [d(i) - \mathbf{u}_i^T \cdot \mathbf{w}_{i-1}] \,. \tag{20.5}$$

Subtracting the optimal coefficient vector $\mathbf{w}$ from both sides and squaring the resulting expressions, we obtain

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

Expanding the right-hand side of this relationship and rearranging terms leads to the equality

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

The right-hand side in the above equality is the product of three terms. Two of these terms, $\mu(i)$ and $|e_a(i) + v(i)|^2$, are nonnegative, whereas the term $(\mu(i) \cdot \|\mathbf{u}_i\|^2 - 1)$ can be positive, negative, or zero depending on the relative magnitudes of $\mu(i)$ and $\|\mathbf{u}_i\|^2$. If we define $\bar{\mu}(i)$ as (assuming nonzero regression vectors):

$$\bar{\mu}(i) = \|\mathbf{u}_i\|^{-2} \,, \tag{20.6}$$

then the following relations hold:

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

The result for $0 < \mu(i) \leq \bar{\mu}(i)$ has a nice interpretation. It states that, no matter what the value of $v(i)$ is and no matter how far $\mathbf{w}_{i-1}$ is from $\mathbf{w}$, the sum of the two energies $\|\tilde{\mathbf{w}}_i\|^2 + \mu(i) \cdot |e_a(i)|^2$ will always be smaller than or equal to the sum of the two disturbance energies $\|\tilde{\mathbf{w}}_{i-1}\|^2 + \mu(i) \cdot |v(i)|^2$. This relationship is a statement of the *passivity* of the algorithm locally in time, as it holds for every time instant. Similar relationships can be developed in terms of the *a posteriori* estimation error.

Since this relationship holds for each time instant i, it also holds over an interval of time such that

$$\frac{\|\tilde{\mathbf{w}}_N\|^2 \; + \; \sum_{i=0}^{N} |\bar{e}_a(i)|^2}{\|\tilde{\mathbf{w}}_{-1}\|^2 + \sum_{i=0}^{N} |\bar{v}(i)|^2} \;\leq\; 1 \,, \tag{20.7}$$

where we have introduced the normalized *a priori* residuals and noise signals

$$\bar{e}_a(i) = \sqrt{\mu(i)} \; e_a(i) \qquad \text{and} \qquad \bar{v}(i) = \sqrt{\mu(i)} \; v(i) \,,$$

respectively. Equation (20.7) states that the lower-triangular matrix that maps the normalized noise signals $\{\bar{v}(i)\}_{i=0}^{N}$ and the initial uncertainty $\tilde{\mathbf{w}}_{-1}$ to the normalized *a priori* residuals $\{\bar{e}_a(i)\}_{i=0}^{N}$ and the final weight error $\tilde{\mathbf{w}}_N$ has a maximum singular value that is less than one. Thus, it is a *contraction mapping* for $0 < \mu(i) \leq \bar{\mu}(i)$. For the special case of a constant step-size μ, this is the same mapping $\mathcal{T}$ that we introduced earlier (20.4).

In the above derivation, we have assumed for simplicity of presentation that the denominators of all expressions are nonzero. We can avoid this restriction by working with differences rather than ratios. Let $\Delta_N(\mathbf{w}_{-1}, v(\cdot))$ denote the difference between the numerator and the denominator of (20.7), such that

$$\Delta_N(\mathbf{w}_{-1}, v(\cdot)) = \left\{ \|\tilde{\mathbf{w}}_N\|^2 + \sum_{i=0}^{N} |\bar{e}_a(i)|^2 \right\} - \left\{ \|\tilde{\mathbf{w}}_{-1}\|^2 + \sum_{i=0}^{N} |\bar{v}(i)|^2 \right\} . \tag{20.8}$$

Then, a similar argument that produced (20.7) can be used to show that for any $\{\mathbf{w}_{-1}, v(\cdot)\}$,

$$\Delta_N(\mathbf{w}_{-1}, v(\cdot)) \leq 0 . \tag{20.9}$$

20.7 Min-Max Optimality of Adaptive Gradient Algorithms

The property in (20.7) or (20.9) is valid for any initial guess $\mathbf{w}_{-1}$ and for any noise sequence $v(\cdot)$, so long as the $\mu(i)$ are properly bounded by $\bar{\mu}(i)$. One might then wonder whether the bound in (20.7) is tight or not. In other words, are there choices $\{\mathbf{w}_{-1}, v(\cdot)\}$ for which the ratio in (20.7) can be made arbitrarily close to one or Δ_N in (20.9) arbitrarily close to zero? We now show that there are. We can rewrite the gradient recursion of (20.5) in the equivalent form

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu(i) \cdot \mathbf{u}_i \cdot [e_a(i) + v(i)] . \tag{20.10}$$

Envision a noise sequence $v(i)$ that satisfies $v(i) = -e_a(i)$ at each time instant i. Such a sequence may seem unrealistic but is entirely within the realm of our unrestricted model of the unknown disturbances. In this case, the above gradient recursion trivializes to $\mathbf{w}_i = \mathbf{w}_{i-1}$ for all i, thus leading to $\mathbf{w}_N = \mathbf{w}_{-1}$. Thus, Δ_N in (20.8) will be zero for this particular experiment. Therefore,

$$\max_{\{\mathbf{w}_{-1}, v(\cdot)\}} \{\Delta_N(\mathbf{w}_{-1}, v(\cdot))\} = 0 .$$

We now consider the following question: how does the gradient recursion in (20.5) compare with other possible causal recursive algorithms for the update of the weight estimate? Let $\mathcal{A}$ denote any given causal algorithm. Suppose that we initialize algorithm $\mathcal{A}$ with $\mathbf{w}_{-1} = \mathbf{w}$, and suppose the noise sequence is given by $v(i) = -e_a(i)$ for $0 \leq i \leq N$. Then, we have

$$\sum_{i=0}^{N} |\bar{v}(i)|^2 = \sum_{i=0}^{N} |\bar{e}_a(i)|^2 \;\leq\; \|\tilde{\mathbf{w}}_N\|^2 + \sum_{i=0}^{N} |\bar{e}_a(i)|^2 ,$$

no matter what the value of $\tilde{\mathbf{w}}_N$ is. This particular choice of initial guess ($\mathbf{w}_{-1} = \mathbf{w}$) and noise sequence $\{v(\cdot)\}$ will always result in a nonnegative value of Δ_N in (20.8), implying for any causal algorithm $\mathcal{A}$ that

$$\max_{\{\mathbf{w}_{-1}, v(\cdot)\}} \{\Delta_N(\mathbf{w}_{-1}, v(\cdot))\} \geq 0 .$$

For the gradient recursion in (20.5), the maximum has to be exactly zero because the global property (20.9) provided us with an inequality in the other direction. Therefore, the algorithm in (20.5) solves the following optimization problem:

$$\min_{\textit{Algorithm}} \left\{ \max_{\{\mathbf{w}_{-1}, v(\cdot)\}} \Delta_N(\mathbf{w}_{-1}, v(\cdot)) \right\} ,$$

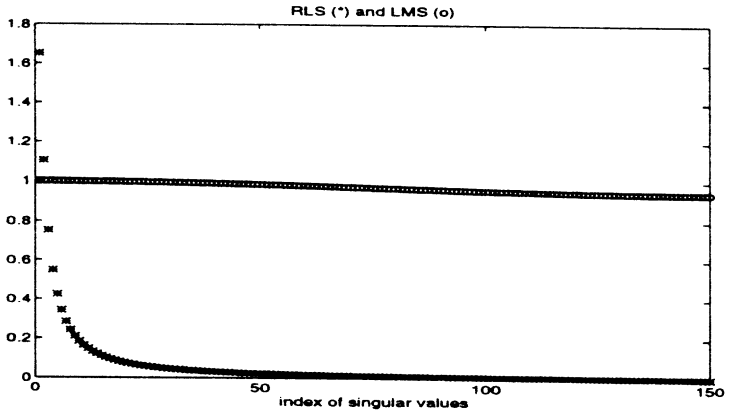


FIGURE 20.3: Singular value plot.

and the optimal value is equal to zero. More details and justification can be found in the references at the end of this chapter, especially connections with so-called H_∞ estimation theory.

As explained before, Δ_N measures the difference between the output energy and the input energy of the algorithm mapping $\mathcal{T}$. The gradient algorithm in (20.5) minimizes the maximum possible difference between these two energies over all disturbances with finite energy. In other words, it minimizes the effect that the worst-possible input disturbances can have on the resulting estimation-error energy.

20.8 Comparison of LMS and RLS Algorithms

To illustrate the ideas in our discussion, we compare the robustness performance of two classical algorithms: the LMS algorithm (20.2) and the recursive least-squares (RLS) algorithm. More details on the example given below can be found in the reference section at the end of the chapter.

Consider the data model in (20.1) where $\mathbf{u}_i$ is a scalar that randomly assumes the values $+1$ and -1 with equal probability. Let $\mathbf{w} = 0.25$, and let $v(i)$ be an uncorrelated Gaussian noise sequence with unit variance. We first employ the LMS recursion in (20.2) and compute the initial 150 estimates $\mathbf{w}_i$, starting with $\mathbf{w}_{-1} = 0$ and using $\mu = 0.97$. Note that μ satisfies the requirement $\mu \leq 1/\|\mathbf{u}_i\|^2 = 1$ for all i. We then evaluate the entries of the resulting mapping $\mathcal{T}$, now denoted by $\mathcal{T}_{lms}$, that we defined in (20.4). We then compute the corresponding $\mathcal{T}_{rls}$ for the recursive-least-squares (RLS) algorithm for these signals, which for this special data model can be expressed as

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \frac{p_i \mathbf{u}_i}{1 + p_i}[d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1}] \ , \quad p_{i+1} = \frac{p_i}{1 + p_i} \ .$$

The initial condition chosen for p_i is $p_0 = \mu = 0.97$.

Figure 20.3 shows a plot of the 150 singular values of the resulting mappings $\mathcal{T}_{lms}$ and $\mathcal{T}_{rls}$. As predicted from our analysis, the singular values of $\mathcal{T}_{lms}$, indicated by an almost horizontal line at unity, are all bounded by one, whereas the maximum singular value of $\mathcal{T}_{rls}$ is approximately 1.65. This result indicates that the LMS algorithm is indeed more robust than the RLS algorithm, as is predicted by the earlier analysis.

Observe, however, that most of the singular values of $\mathcal{T}_{rls}$ are considerably smaller than one, whereas the singular values of $\mathcal{T}_{lms}$ are clustered around one. This has an interesting interpretation that we explain as follows. An $N \times N$-dimensional matrix A has N singular values $\{\sigma_i\}$ that are equal

to the positive square-roots of the eigenvalues of AA^T. For each σ_i, there exists a unit-norm vector x_i such that the energy gain from x_i to Ax_i is equal to σ_i^2, i.e., $\sigma_i = \|Ax_i\|/\|x_i\|$. The vector x_i can be chosen as the ith right singular vector of A. Now, recall that $\mathcal{T}_{lms}$ and $\mathcal{T}_{rls}$ are finite-dimensional matrices that map a disturbance vector dist to the estimation-errors vector error. Considering the plot of the singular values of $\mathcal{T}_{rls}$, we see that if the disturbance vector dist happens to lie in the range space of the right singular vectors associated with the smaller singular values in this plot, then its effect will be significantly attenuated. This fact indicates that while the performance of the LMS algorithm guards against worst-case disturbances, the RLS algorithm is likely to have a better performance than the LMS algorithm on average, as is well-known.

20.9 Time-Domain Feedback Analysis

Robust adaptive filters are designed to induce contractive mappings between sequences of numbers. This fact also has important implications on the convergence performance of a robust adaptive scheme. In the remaining sections of this chapter, we discuss the combined issues of robustness and convergence from a deterministic standpoint. In particular, the following issues are discussed:

- We show that each step of the update equation of the gradient algorithm in (20.5) can be described in terms of an elementary section that possesses a useful feedback structure.
- The feedback structure provides insights into the robust and convergence performance of the adaptive scheme. This is achieved by studying the energy flow through a cascade of elementary sections and by invoking a useful tool from system theory known as the *small gain theorem*.
- The feedback analysis extends to more general update relations. The example considered here is filtered-error LMS algorithm, although the methodology can be extended to other structures such as perceptrons. Details can be found in the references at the end of this chapter.

20.9.1 Time-Domain Analysis

From the update equation in (20.5), $\tilde{\mathbf{w}}_i$ satisfies

$$\tilde{\mathbf{w}}_i = \tilde{\mathbf{w}}_{i-1} - \mu(i) \cdot \mathbf{u}_i \cdot [e_a(i) + v(i)] \,. \tag{20.11}$$

If we multiply both sides of (20.11) by $\mathbf{u}_i^T$ from the left, we obtain the following relation among $\{e_p(i), e_a(i), v(i)\}$:

$$e_p(i) = \left(1 - \frac{\mu(i)}{\bar{\mu}(i)}\right) e_a(i) - \frac{\mu(i)}{\bar{\mu}(i)} v(i) \,, \tag{20.12}$$

where $\bar{\mu}(i)$ is given by (20.6). Using (20.12), (20.5) can be rewritten in the equivalent form

$$\begin{aligned} \mathbf{w}_i &= \mathbf{w}_{i-1} + \bar{\mu}(i) \cdot \mathbf{u}_i \cdot [e_a(i) - e_p(i)] \,, \\ &= \mathbf{w}_{i-1} + \bar{\mu}(i) \cdot \mathbf{u}_i \cdot [e_a(i) + r(i)] \,, \end{aligned} \tag{20.13}$$

where we have defined the signal $r(i) = -e_p(i)$ for convenience. The expression (20.13) shows that (20.5) can be rewritten in terms of a new step-size $\bar{\mu}(i)$ and a modified "noise" term $r(i)$.

Therefore, if we follow arguments similar to those prior to (20.6), we readily conclude that for algorithm (20.5) the following equality holds for *all* $\{\mu(i), v(i)\}$:

$$\frac{\|\tilde{\mathbf{w}}_i\|^2 + \bar{\mu}(i)\,|e_a(i)|^2}{\|\tilde{\mathbf{w}}_{i-1}\|^2 + \bar{\mu}(i)\,|r(i)|^2} = 1 \,. \tag{20.14}$$

This relation establishes a lossless map (denoted by $\overline{\mathcal{T}}_i$) from the signals $\{\tilde{\mathbf{w}}_{i-1}, \sqrt{\bar{\mu}(i)}\ r(i)\}$ to the signals $\{\tilde{\mathbf{w}}_i, \sqrt{\bar{\mu}(i)}\ e_a(i)\}$. Correspondingly, using (20.12), the map from the original weighted disturbance $\sqrt{\bar{\mu}(i)}v(i)$ to the weighted estimation error signal $\sqrt{\bar{\mu}(i)}e_a(i)$ can be expressed in terms of the feedback structure shown in Fig. 20.4.

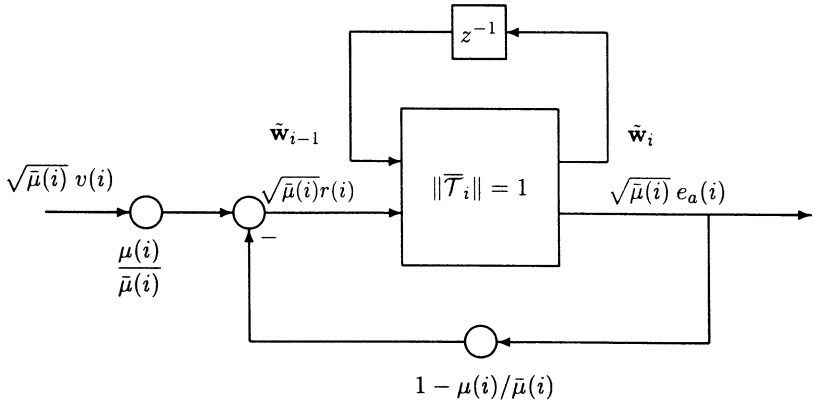


FIGURE 20.4: A time-variant lossless mapping with gain feedback for gradient algorithms. © IEEE 1996. (*Source:* Rupp, M. and Sayed, A.H., A time domain feedback analysis of filtered-error adaptive gradient algorithms, *IEEE Trans. Signal Process.* 44(6): 1428–1439, June 1996. With permission.)

The feedback description provides useful insights into the behavior of the adaptive scheme. Because the map $\overline{\mathcal{T}}_i$ in the feedforward path is lossless or energy-preserving, the design and analysis effort can be concentrated on the terms contained in the feedback path. This feedback block controls:

- How much energy is fed back into the input of each section and whether energy magnification or demagnification may occur (i.e., stability).
- How sensitive the estimation error is to noise and disturbances (i.e., robustness).
- How fast the estimation error energy decays (i.e., convergence rate).

20.9.2 l_2—Stability and the Small Gain Condition

We start by reconsidering the robustness issue. Recall that if the step-sizes $\mu(i)$ are chosen such that $\mu(i) \leq \bar{\mu}(i)$, then robustness is guaranteed in that the ratio of the energies of the signals in (20.7) will be bounded by one.

The condition on $\mu(i)$ can be relaxed at the expense of guaranteeing energy ratios that are bounded by some other positive number, say

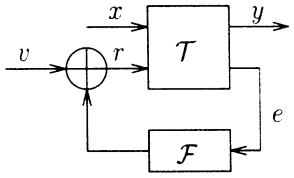
$$\frac{\text{weighted estimation error energy}}{\text{weighted disturbance energy}} \leq \gamma^2\,, \tag{20.15}$$

for some constant γ to be determined. This is still a desirable property because it means that the disturbance energy will be, at most, scaled by a factor of γ. This fact can in turn lead to useful convergence conclusions, as argued later.

In order to guarantee robustness conditions according to (20.15), for some γ, we rely on the observation that feedback configurations of the form shown in Fig. 20.4 can be analyzed using a

tool known in system theory as the small gain theorem. In loose terms, this theorem states that the stability of a feedback configuration such as that in Fig. 20.4 is guaranteed if the product of the norms of the feedforward and the feedback mappings are strictly bounded by one. Since the feedforward mapping $\overline{\mathcal{T}}_i$ has a norm (or maximum singular value) of one, the norm of the feedback map needs to be strictly bounded by one for stability of this system.

To illustrate these concepts more fully, consider the feedback structure in Fig. 20.5 that has a lossless mapping $\mathcal{T}$ in its feedforward path and an arbitrary mapping $\mathcal{F}$ in its feedback path. The input/output signals of interest are denoted by $\{x, y, r, v, e\}$. In this system, the signals x, v play the role of the disturbances.



$$\begin{bmatrix} y \\ e \end{bmatrix} = \mathcal{T} \begin{bmatrix} x \\ r \end{bmatrix}, \quad \|\mathcal{T}\| = 1$$

$$r = v + \mathcal{F}e$$

FIGURE 20.5: A feedback structure.

The losslessness of the feedforward path implies conservation of energy such that

$$\|y\|^2 + \|e\|^2 = \|x\|^2 + \|r\|^2 .$$

Consequently, $\|e\| \leq \|x\| + \|r\|$. On the other hand, the triangle inequality of norms implies that

$$\|r\| \ \leq \ \|v\| + \|\mathcal{F}\| \cdot \|e\| ,$$

where the notation $\|\mathcal{F}\|$ denotes the maximum singular value of the mapping $\mathcal{F}$. Provided that the small gain condition $\|\mathcal{F}\| < 1$ is satisfied, we have

$$\|e\| \ \leq \ \frac{1}{1 - \|\mathcal{F}\|} \cdot [\ \|x\| + \|v\|\ ] . \tag{20.16}$$

Thus, a contractive $\mathcal{F}$ guarantees a robust map from $\{x, v\}$ to $\{e\}$ with a robustness level that is determined by the factor $1/(1 - \|\mathcal{F}\|)$. In this case, we shall say that the map from $\{x, v\}$ to $\{e\}$ is l_2-stable.

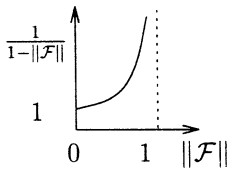


FIGURE 20.6: Plot of the l_2-gain.

A plot of the factor $1/(1 - \|\mathcal{F}\|)$, as a function of $\|\mathcal{F}\|$, is shown in Fig. 20.6. It can be seen that the smaller the value of $\|\mathcal{F}\|$:

- The smaller the effect of $\{x, v\}$ on $\{e\}$.
- The smaller the upper bound on $\|e\|$.

Moreover, we shall argue that smaller values of $\|\mathcal{F}\|$ are associated with faster convergence. Therefore, controlling the norm of $\mathcal{F}$, is important for both the robustness and convergence performance of an adaptive algorithm. In most cases, the feedback filter $\mathcal{F}$ will depend on several quantities, such as the step-sizes $\{\mu(i)\}$ and the data vectors $\{\mathbf{u}_i\}$ (as in Fig. 20.4). It may also depend on error filters and on regression filters that appear in more general adaptive schemes.

Referring to Fig. 20.4, define

$$\eta(N) = \max_{0\le i\le N} \left|1 - \frac{\mu(i)}{\bar{\mu}(i)}\right| \quad \text{and} \quad \xi(N) = \max_{0\le i\le N} \frac{\mu(i)}{\bar{\mu}(i)}\,.$$

That is, $\eta(N)$ is the maximum absolute value of the gain of the feedback loop over the interval of time $0 \le i \le N$, and $\xi(N)$ is the maximum value of the scaling factor $\mu(i)/\bar{\mu}(i)$ at the input of the feedback interconnection. In this context, the small gain condition requires that $\eta(N) < 1$. This condition is equivalent to choosing the step-size parameter $\mu(i)$ such that $0 < \mu(i) < 2\bar{\mu}(i)$. Under this condition, the general relation (20.16) can be used to deduce either of the following two relationships:

$$\sqrt{\sum_{i=0}^{N} \bar{\mu}(i)\,|e_a(i)|^2} \le \frac{1}{1-\eta(N)} \left[\|\tilde{\mathbf{w}}_{-1}\| + \xi(N) \sqrt{\sum_{i=0}^{N} \bar{\mu}(i)|v(i)|^2} \right] \tag{20.17}$$

or

$$\sqrt{\sum_{i=0}^{N} \mu(i)\,|e_a(i)|^2} \le \frac{\xi^{1/2}(N)}{1-\eta(N)} \left[\|\tilde{\mathbf{w}}_{-1}\| + \xi^{1/2}(N) \sqrt{\sum_{i=0}^{N} \mu(i)|v(i)|^2} \right]. \tag{20.18}$$

Note that in either case the upper bound on $\mu(i)$ is now $2\bar{\mu}(i)$ and the robustness level is essentially determined by

$$\frac{1}{1-\eta(N)} \quad \text{or} \quad \frac{\xi^{1/2}(N)}{1-\eta(N)}\,,$$

depending on how the estimation errors $\{e_a(i)\}$ and the noise terms $\{v(i)\}$ are normalized [by $\mu(\cdot)$ or $\bar{\mu}(\cdot)$].

20.9.3 Energy Propagation in the Feedback Cascade

By studying the energy flow in the feedback interconnection of Fig. 20.4, we can also obtain some physical insights into the convergence behavior of the gradient recursion (20.5).

Assume that $\mu(i) = \bar{\mu}(i)$, such that the feedback loop of Fig. 20.4 is disconnected. In this situation, there is no energy flowing back into the lower input of the lossless section from its lower output $e_a(\cdot)$. The losslessness of the feedforward path then implies that

$$E_w(i) = E_w(i-1) + E_v(i) - E_e(i),$$

where we have defined the energy terms

$$E_e(i) = \bar{\mu}(i)|e_a(i)|^2, \quad E_v(i) = \bar{\mu}(i)|v(i)|^2, \quad E_w(i) = \|\tilde{\mathbf{w}}_i\|^2\,.$$

In the noiseless case where $v(i) = 0$, the above expression implies that the weight-error energy is a nonincreasing function of time, i.e., $E_w(i) \le E_w(i-1)$.

However, what happens if $\mu(i) \neq \bar{\mu}(i)$? In this case, the feedback path is active and the convergence speed will be affected because the rate of decrease in the energy of the estimation error will change. Indeed, for $\mu(i) \neq \bar{\mu}(i)$ we obtain for $E_v(i) = 0$

$$E_w(i) = E_w(i-1) - \left(1 - \left|1 - \frac{\mu(i)}{\bar{\mu}(i)}\right|^2\right) E_e(i) \,,$$

where, due to the small gain condition, the coefficient multiplying $E_e(i)$ can be seen to be smaller than 1.

Loosely speaking, this energy argument indicates for $v(i) = 0$ that the smaller the maximum singular value of feedback block $\mathcal{F}$, for a generic feedback interconnection of the form shown in Fig. 20.5, the faster the convergence of the algorithm will be, since less energy is fed back to the input of each section.

20.9.4 A Deterministic Convergence Analysis

The energy argument can be pursued in order to provide sufficient deterministic conditions for the convergence of the weight estimates $\mathbf{w}_i$ to the true weight vector $\mathbf{w}$. The argument follows as a consequence of the energy relations (or robustness bounds) (20.17) and (20.18), which essentially establishes that the adaptive gradient algorithm (20.5) maps a finite-energy sequence to another finite-energy sequence.

To clarify this point, we define the quantities

$$\eta = \sup_i \left|1 - \frac{\mu(i)}{\bar{\mu}(i)}\right|, \quad \xi = \sup_i \left[\frac{\mu(i)}{\bar{\mu}(i)}\right] \,,$$

and note that if the step-size parameter $\mu(i)$ is chosen such that $\mu(i)\|\mathbf{u}_i\|^2$ is uniformly bounded by 2, then we guarantee $\xi < 2$ and $\eta < 1$. We further note that it follows from the weight-error update relation (20.11) that $\tilde{\mathbf{w}}_i$ satisfies

$$\tilde{\mathbf{w}}_i = \tilde{\mathbf{w}}_{i-1} - \bar{\mathbf{u}}_i^T [\bar{e}_a(i) + \bar{v}(i)] \,, \tag{20.19}$$

where we have defined $\bar{\mathbf{u}}_i = \sqrt{\mu(i)}\, \mathbf{u}_i$ [likewise for $\bar{e}_a(i)$, $\bar{v}(i)$]. The following conclusions can now be established under the stated conditions.

- *Finite noise energy condition.* We assume that the normalized sequence $\{\bar{v}(\cdot) = \sqrt{\mu(i)}\, v(i)\}$ has finite energy, i.e.,

$$\sum_{i=0}^{\infty} \mu(i)|v(i)|^2 < \infty \,. \tag{20.20}$$

  This in turn implies that $\bar{v}(i) \to 0$ as $i \to \infty$ (but not necessarily $v(i) \to 0$). If the initial weight-error vector is finite, $\|\tilde{\mathbf{w}}_{-1}\| < \infty$, then condition (20.20) along with the energy bound (20.18) [as $N \to \infty$] allows us to conclude that $\sum_{i=0}^{\infty} \mu(i)|e_a(i)|^2 < \infty$. Consequently, $\lim_{i\to\infty} \bar{e}_a(i) \to 0$ (but not necessarily $e_a(i) \to 0$).

- *Persistent excitation condition.* We also assume that the normalized vectors $\{\bar{\mathbf{u}}_i\}$ are persistently exciting. By this we mean that there exists a finite integer $L \geq M$ such that the smallest singular value of

$$\text{col}\,\{\bar{\mathbf{u}}_i^T, \bar{\mathbf{u}}_{i+1}^T, \ldots, \bar{\mathbf{u}}_{i+L}^T\}$$

  is uniformly bounded from below by a positive quantity for sufficiently large i. The persistence of excitation condition can be used to further conclude from $\bar{e}_a(i) \to 0$ that $\lim_{i\to\infty} \mathbf{w}_i = \mathbf{w}$.

The above statements can also be used to clarify the behavior of the adaptive algorithm (20.5) in the presence of finite-power (rather than finite-energy) normalized noise sequences $\{\bar{v}(\cdot)\}$, i.e., for $v(\cdot)$ satisfying

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \mu(i)|v(i)|^2 = P_v \ < \infty \ .$$

For this purpose, we divide both sides of (20.18) by $\sqrt{N}$ and take the limit as $N \to \infty$ to conclude that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \mu(i)|e_a(i)|^2 \leq \frac{\xi^2 P_v}{(1-\eta)^2} \ .$$

In other words, a bounded noise power leads to a bounded estimation error power.

20.10 Filtered-Error Gradient Algorithms

The feedback analysis of the former sections can be extended to gradient algorithms that employ filtered versions of the error signal $d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1}$. Such algorithms are useful in applications such as active noise and vibration control and in adaptive IIR filters, where a filtered error signal is more easily observed or measured. Figure 20.7 depicts the context of this problem. The symbol F denotes the filter that operates on $e(i)$. For our discussion, we assume that F is a finite-impulse response filter of order M_F, such that the z-transform of its impulse response is $F(z) = \sum_{j=0}^{M_F-1} f_j z^{-j}$.

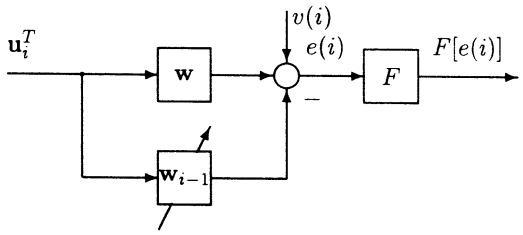


FIGURE 20.7: Structure of filtered-error gradient algorithms. © IEEE 1996. (*Source:* Rupp, M. and Sayed, A.H., A time domain feedback analysis of filtered-error adaptive gradient algorithms, *IEEE Trans. Signal Process.* 44(6): 1428–1439, June 1996. With permission.)

For purposes of discussion, we focus on one particular form of adaptive update known as the *filtered-error LMS algorithm:*

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \mu(i) \cdot \mathbf{u}_i \cdot F[d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1}] \ . \tag{20.21}$$

Comparing (20.21) with (20.5), the only difference between the two updates is the filter F that acts on the error $d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1}$.

Following the discussion that led to (20.13), it can be verified that (20.21) is equivalent to the following update:

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \bar{\mu}(i) \cdot \mathbf{u}_i \cdot [e_a(i) + r(i)] \ , \tag{20.22}$$

where $\bar{\mu}(i) = 1/\|\mathbf{u}_i\|^2$, $e_a(i) = \mathbf{u}_i^T \tilde{\mathbf{w}}_{i-1}$, and $r(i)$ is defined as

$$\bar{\mu}(i) r(i) = \mu(i) F[v(i)] - \bar{\mu}(i) e_a(i) + \mu(i) F[e_a(i)] \,. \tag{20.23}$$

Expression (20.22) is of the same form as (20.13), which implies that the following relation also holds:

$$\frac{\|\tilde{\mathbf{w}}_i\|^2 + \bar{\mu}(i)\,|e_a(i)|^2}{\|\tilde{\mathbf{w}}_{i-1}\|^2 + \bar{\mu}(i)\,|r(i)|^2} = 1 \,. \tag{20.24}$$

This establishes that the map $\overline{\mathcal{T}}_i$ from the signals $\{\tilde{\mathbf{w}}_{i-1}, \sqrt{\bar{\mu}(i)} r(i)\}$ to the signals $\{\tilde{\mathbf{w}}_i, \sqrt{\bar{\mu}(i)} e_a(i)\}$ is lossless. Moreover, the map from the original disturbance $\sqrt{\bar{\mu}(\cdot)} v(\cdot)$ to the signal $\sqrt{\bar{\mu}(\cdot)} e_a(\cdot)$ can be expressed in terms of a feedback structure, as shown in Fig. 20.8. We remark that the notation $1 - \frac{\mu(i)}{\sqrt{\bar{\mu}(i)}}\, F[\cdot]\, \frac{1}{\sqrt{\bar{\mu}(i)}}$ should be interpreted as follows. We first divide $\sqrt{\bar{\mu}(i)}\; e_a(i)$ by $\sqrt{\bar{\mu}(i)}$ before filtering it by the filter F and then scaling the result by $\mu(i)/\sqrt{\bar{\mu}(i)}$. Similarly, the term $\sqrt{\bar{\mu}(i)}\; v(i)$ is first divided by $\sqrt{\bar{\mu}(i)}$, then filtered by F, and is finally scaled by $\mu(i)/\sqrt{\bar{\mu}(i)}$.

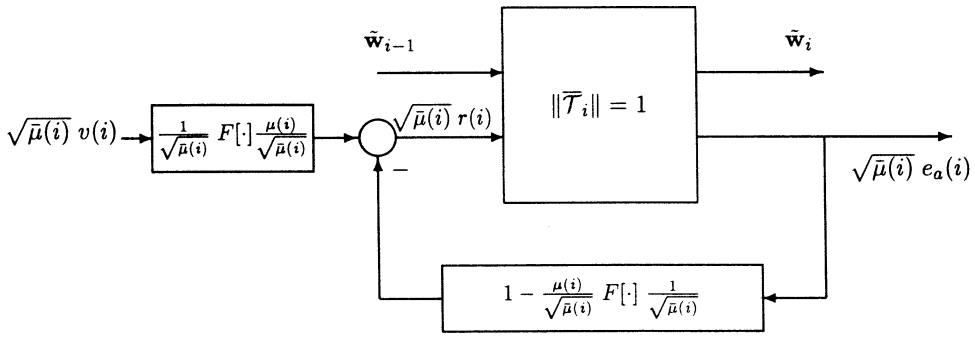


FIGURE 20.8: Filtered-error LMS algorithm as a time-variant lossless mapping with dynamic feedback. © IEEE 1996. (*Source:* Rupp, M. and Sayed, A.H., A time domain feedback analysis of filtered-error adaptive gradient algorithms, *IEEE Trans. Signal Process.* 44(6): 1428–1439, June 1996. With permission.)

The feedback path now contains a dynamic system. The small gain theorem dictates that this system will be robust if the feedback path is a contractive system. For the special case of the *projection* filtered-error LMS algorithm that employs the step-size $\mu(i) = \alpha\; \bar{\mu}(i),\; \alpha > 0$,

$$\mathbf{w}_i = \mathbf{w}_{i-1} + \alpha \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|^2}\; F[d(i) - \mathbf{u}_i^T \mathbf{w}_{i-1}] \,, \tag{20.25}$$

the small-gain condition implies that the following matrix should be strictly contractive:

$$\mathbf{P}_N = \begin{pmatrix} 1 - \alpha f_0 & & \mathbf{O} & \\ -\alpha \frac{\sqrt{\bar{\mu}(1)}}{\sqrt{\bar{\mu}(0)}} f_1 & 1 - \alpha f_0 & & \\ -\alpha \frac{\sqrt{\bar{\mu}(2)}}{\sqrt{\bar{\mu}(0)}} f_2 & -\alpha \frac{\sqrt{\bar{\mu}(2)}}{\sqrt{\bar{\mu}(1)}} f_1 & 1 - \alpha f_0 & \\ \vdots & & & \ddots \end{pmatrix} .$$

Here, the $\{f_i\}$ are the coefficients of the FIR filter F. Since, in practice, the length M_F of this filter is usually much smaller than the length of the regression vector $\mathbf{u}_i$, the energy of the input sequence $\mathbf{u}_i$

does not change very rapidly over the filter length M_F, such that

$$\bar{\mu}(i) \approx \ldots \approx \bar{\mu}(i - M_F) \ .$$

In this case, $\mathbf{P}_N$ becomes

$$\mathbf{P}_N \approx \mathbf{I} - \alpha \mathbf{F}_N \ , \tag{20.26}$$

where $\mathbf{F}_N$ is the lower triangular Toeplitz matrix that describes the convolution of the filter F on an input sequence. This is generally a banded matrix since $M_F \ll M$, as shown below for the special case of $M_F = 3$,

$$\mathbf{F}_N = \begin{bmatrix} f_0 & & & & \\ f_1 & f_0 & & & \\ f_2 & f_1 & f_0 & & \\ & f_2 & f_1 & f_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} .$$

In this case, the strict contractivity of $(\mathbf{I} - \alpha \mathbf{F}_N)$ can be guaranteed by choosing the step-size parameter α such that

$$\max_{\Omega} \left| 1 - \alpha F(e^{j\Omega}) \right| < 1 \ , \tag{20.27}$$

where $F(z)$ is the transfer function of the error filter. For better convergence performance, we may choose α by solving the min-max problem

$$\min_{\alpha} \ \max_{\Omega} \left| 1 - \alpha F(e^{j\Omega}) \right| \ . \tag{20.28}$$

If the resulting minimum is less than one, then the corresponding optimum value of α will result in faster convergence, and it will also guarantee the robustness of the scheme.

We now illustrate these concepts via a simulation example. The error-path filter for this example is

$$F(z) = 1 - 1.2\, z^{-1} + 0.72\, z^{-2} \ .$$

We use an FIR filter adapted by the algorithm in (20.25), where the input signal to the adaptive filter consists of a single sinusoid of frequency $\Omega_0 = 1.2/\pi$. In this case, if we assume that the *a priori* error signal is dominated by the frequency component Ω_0, we can solve for the optimum α via the simpler expression [cf. (20.28)] $\min_\alpha \left| 1 - \alpha F\left(e^{j\Omega_o}\right) \right|$. The resulting optimum value of α is

$$\alpha_{opt} = \text{Real} \left\{ \frac{1}{F(e^{-j\Omega_0})} \right\} \ .$$

This step size provides the fastest convergence speed. In addition, the stability limits for α can be shown to be $0 < \alpha < 2\alpha_{opt}$ using a similar procedure.

Figure 20.9 shows three convergence curves of the average squared error $\text{Av}[|e(i)|^2] = \frac{1}{50} \sum_{j=1}^{50} |e_j(i)|^2$, as determined from 50 simulation runs of the projection filtered-error LMS algorithm for the choices $\alpha = 0.085$, $\alpha = 0.15$ and $\alpha = 0.18$, respectively. In this case, we have generated an input sequence of the form $u(i) = \sin[1.2i + \phi]$ for each simulation run, where ϕ is uniformly chosen from the interval $[-\pi, \pi]$ to obtain smoother learning curves after averaging, and the $M = 10$ coefficients of the unknown system were all set to unity. Moreover, the additive noise $v(i)$ corrupting the signal $d(i)$ is uncorrelated Gaussian-distributed with a level that is -40dB below that of the input signal power. The optimal step-size α_{opt} in this case can be calculated to be $\alpha_{opt} = 0.085$ and the stability bounds for the system are $0 < \alpha < 0.17$. As expected, choosing $\alpha = 0.085$ provides the fastest convergence speed for this situation. We also see that for values of α greater than $2\alpha_{opt}$, the error of the system diverges.

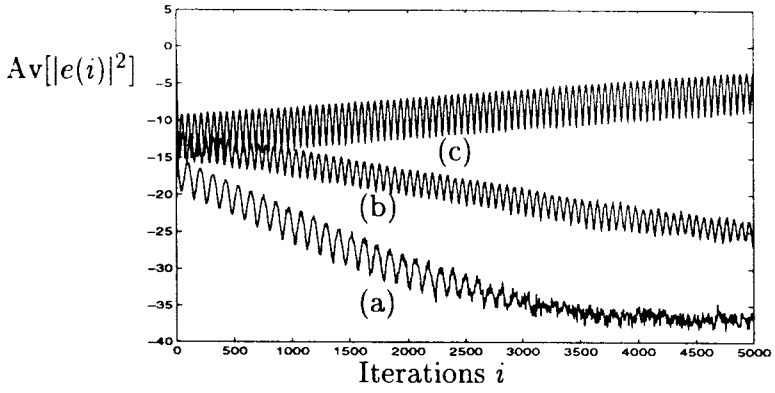


FIGURE 20.9: Convergence behavior for FELMS algorithm with sinusoidal input sequence and various step-sizes $\alpha = 0.085$(a), 0.15 (b), 0.18(c). © IEEE 1996. (*Source:* Rupp, M. and Sayed, A.H., A time domain feedback analysis of filtered-error adaptive gradient algorithms, *IEEE Trans. Signal Process.* 44(6): 1428–1439, June 1996. With permission.)

20.11 References and Concluding Remarks

The intent of this chapter was to highlight certain robustness and convergence issues that arise in the study of adaptive algorithms in the presence of uncertain data. More details, extensions, and related discussions can be found in several of the references indicated in this section. The references are not intended to be complete but rather indicative of the work in the different areas. More complete lists can be found in several of the textbooks mentioned herein.

Detailed discussions on the different forms of adaptive algorithms and their potential applications can be found in:

[1] Haykin, S., *Adaptive Filter Theory*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.

[2] Proakis, J.G., Rader, C.M., Ling, F., and Nikias, C.L., *Advanced Digital Signal Processing*, Macmillan Publishing, New York, 1992.

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

[4] Sayed, A.H. and Kailath, T., A state-space approach to adaptive RLS filtering, *IEEE Signal Processing Magazine*, 11(3), 18–60, July 1994.

The fundamentals of robust or H^∞ design, both in filtering and control applications, can be found in the following references:

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

[6] Zhou, K., Doyle, J.C., and Glover, K., *Robust and Optimal Control*, Prentice Hall, Englewood Cliffs, NJ, 1996.

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

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

[9] Hassibi, B., Sayed, A.H., and Kailath, T., Linear estimation in Krein spaces — Part I: Theory, *IEEE Trans. Automatic Control*, 41(1), 18–33, Jan. 1996.

[10] Hassibi, B., Sayed, A.H., and Kailath, T., Linear estimation in Krein spaces — Part II: Applications, *IEEE Trans. Automatic Control*, 41(1), 34–49, Jan. 1996.

The small gain analysis is a standard tool in linear and nonlinear system theory. More advanced and detailed treatments can be found in:

[11] Khalil, H.K., *Nonlinear Systems*, 2nd ed., Macmillan, New York, 1996.
[12] Vidyasagar, M., *Nonlinear Systems Analysis*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ, 1993.

The LMS algorithm usually has been presented in the literature as an instantaneous-gradient-based approximation for the steepest descent algorithm. Its robustness properties, and the interesting observation that it is in fact the exact solution of a min-max (or H^∞) optimization problem, have been first noted in

[13] Hassibi, B., Sayed, A.H., and Kailath, T., H^∞ optimality of the LMS algorithm, *IEEE Trans. Signal Processing*, 44(2), 267–280, Feb. 1996. See also *Proc. CDC*, 1, 74–79, 1993.

Also, more details on the example comparing the performance of LMS and RLS can be found in the above reference. Extensions of the discussion to the backpropagation algorithm for neural network training, and other related results in adaptive filtering and H_∞ estimation and control can be found in

[14] Hassibi, B., Sayed, A.H., and Kailath, T., LMS and backpropagation are minimax filters, in *Neural Computation and Learning*, Roychowdhurys, V., Siu, K.Y., and Orlitsky, A., Eds., Kluwer Academic Publishers, 1994, 425–447.
[15] Hassibi, B., *Indefinite Metric Spaces in Estimation, Control, and Adaptive Filtering*, Ph.D. Dissertation, Stanford University, August 1996.
[16] Hassibi, B., Sayed, A.H., and Kailath, T., *Indefinite Quadratic Estimation and Control: A Unified Approach to H_2 and H_∞ Theories*, to be published by SIAM, Studies in Applied Mathematics Series, 1997.

Extensions of the feedback analysis to Perceptron training in neural network can be found in

[17] Rupp, M. and Sayed, A.H., Supervised learning of perceptron and output feedback dynamic networks: a feedback analysis via the small gain theorem, *IEEE Trans. Neural Networks*, 8(3), 612–622, May 1997.

A Cauchy-Schwarz argument that further highlights the robustness property of adaptive gradient algorithms, along with other local energy bounds, are given in

[18] Sayed, A.H. and Rupp, M., Error energy bounds for adaptive gradient algorithms, *IEEE Trans. Signal Processing*, 44(8), 1982–1989, Aug. 1996.
[19] Sayed, A.H. and Kailath, T., A state-space approach to adaptive RLS filtering, *IEEE Signal Processing Magazine*, 11(3), 18–60, July 1994.

The time-domain feedback and small gain analyses of adaptive filters, along with extensions to nonlinear settings and connections with Gauss-Newton updates and H^∞ filters, are discussed in

[20] Rupp, M. and Sayed, A.H., A time-domain feedback analysis of filtered-error adaptive gradient algorithms, *IEEE Trans. on Signal Processing*, 44(6), 1428–1439, June 1996.
[21] Rupp, M. and Sayed, A.H., Robustness of Gauss-Newton recursive methods: a deterministic feedback analysis, *Signal Processing*, 50(3), 165–188, June 1996.
[22] Sayed, A.H. and Rupp, M., An l_2-stable feedback structure for nonlinear adaptive filtering and identification, *Automatica*, 33(1), 13–30, 1997.

Discussions of the singular value decomposition and its properties can be found in

[23] Golub, G.H. and Van Loan, C.F., *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.