Decision Feedback Equalizers: Concepts Towards Design Guidelines

B. D. O. ANDERSON, R. A. KENNEDY and P. R. BITMEAD
Department of Systems Engineering, Research School of Physical Sciences,Australian National University,
G.P.O. Box 4, Canberra A.C.T., 2001, Australia.

* Research supported by the Australian Telecommunications and Electronics Research Board (A.T.E.R.B.)

ABSTRACT: A quantitative and qualitative analysis of the stochastic dynamics of the blind adaptation of decision feedback equalizers (DFEs) operating on noiseless communication channels is presented. The overall goal, not yet achieved, is to identify channel classes for which DFE use is a practical option. Averaging analysis is invoked to predict the existence and location of undesirable attraction points to the adaptation algorithm. The parameter space in which adaptation takes place is shown to be subdivided into a large but finite set of convex polytopes when the channel can be modelled as a finite impulse response filter. With each polytope we show that the parameters describing the averaged equation are fixed. Further, for every point in a given polytope the error performance is described by a single finite state Markov process. Hence, the adaptive dynamics and the error properties of a noiseless DFEs do not depend continuously with the operating point in parameter space but rather only as a function of polytope membership. The possibility arises of the algorithm becoming hung at an inappropriate attraction point.

1. INTRODUCTION

Equalization is the process of cancelling out the distorting effect of a linear channel. Decision feedback equalization refers to two properties of a particular type of equalizer depicted in Fig.1: the cancellation is effectuated by feeding back the equalizer output (which is supposed to be identical to the channel input, perhaps with delay), and a slicer is used to decide whether the equalizer output should be +1 or -1. Of course such an equalizer is only of use when the a_k sequence takes values ±1. (Generalizations to M-level sequences are of course possible.)

In adaptive decision feedback equalizers (DFEs), the d_i taps are adjusted by processing in some way the outputs of the equalizer. When the input sequence a_k giving rise to the outputs is not known, which is common, the equalization is blind. The best setting of the d_i is d_i = h_i (assuming h_i > 0 for the moment); then if a sequence of ak of length N is correctly decoded, equalization will thereafter be perfect, the feedback signal exactly cancelling the distortion. Of course, equalization can be disturbed by a large enough noise pulse on the channel output.

What are the key issues? First, in relation to non-adaptive equalization, one can note that even if d_i = h_i for all i, noise, or incorrect initial conditions can mean that initially, a_0 ≠ a_k. Then the key question arises: how long will it take to obtain correct equalization or recovery from noise? For practical channels, equalization seems rapid. However, one can readily contrive channel impulse responses for which, at typical bit rates and impulse response lengths, correct equalization will only occur after 10^10 years on average [1]. So the real question is: for what classes of channels will equalization occur in an acceptable time? Some partial answers are known, [2]. Incidentally, tools for evaluating the time to equalization reasonably accurately are available [3,4]. However, there is a need for further analytic tools.

Once one turns to adaptive identification schemes, where the d_i are initially different form the h_i, the range of questions is even more broad. The most general questions are those of relating the time of convergence of the adjustable d_i to acceptable values to certain properties of the channel. Going along with this are questions of deriving analytic tools, and identifying classes of channels for which one has an acceptable overall equalization time (resulting from both d_i adjustment and recovery from noise or inappropriate initial conditions).

The adaptation algorithm for adjusting the d_i can be thought of as searching for a minimum; a noisy gradient algorithm is most common. As it turns out, there can be more than one local minimum of the underlying objective function. So we have to be concerned with how long one takes to get to a local minimum, and whether the errors in estimating the gradient, interpreted as a noise disturbance, can cause movement away from one local minimum to another. Again, one issue is whether or not a particular local minimum corresponds to correct equalization or not. This occurrence of multiple local minima has been observed in other forms of equalizer, see e.g., [5-9].

The analysis of some of these questions is pursued in this paper. One of our principal tools is averaging theory, which applies when the time scale governing adjustment of the d_i is much slower than the time scale associated with the channel and non-adaptive DFE stochastic dynamics. The issue of jumping out from one local minimum to reach a better one is hardly addressed. The theoretical tools which will ultimately give typical times are those of large deviation theory including importance sampling [10], and the recent theories of [11,12] developed to study systems excited by small amounts of noise (modelling the gradient estimation errors in our problem).

Much of this paper is a consequence of the ideas presented by Jennings [13] which also addresses the problem of blind adaptation of DFEs.

The paper is structured as follows. In section 2, we provide a precise description of the adaptive DFE. In section 3, we establish that DFE performance depends on the feedback tap weights d_i in a piecewise constant manner. Section 4 discusses the modelling of channel and DFE using finite state Markov processes. Section 5 discusses the equilibria or local minima using averaging theory, and in
section 6 we discuss qualitative features of the tap weight variations. Section 7 contains concluding remarks.

2 SYSTEM DESCRIPTION AND NOTATION

The system we consider is shown in Fig.1. The communication channel will be modeled as a finite impulse response (FIR) filter, with impulse response \( h_0, h_1, \ldots, h_N \), driven by binary data \( x \in \{-1, +1\} \) (k denotes the discrete time index). Channel noise \( n \) is depicted in Fig.1; however, we will regard its influence as secondary to the analysis. Therefore, we make our first assumption:

(A1) Channel noise is negligible, i.e., \( n = 0 \).

The DFE consists of a \( N \)-tap delay line, represented by weights \( d_1, d_2, \ldots, d_N \) adapted to minimize the (residual) intersymbol interference (ISI). This tapped delay line is fed by past decisions or data estimates \( \hat{h}_k \in \{-1, +1\} \) and this can lead to problems when past decisions are incorrect, an effect referred to as error propagation [14]. Note that in Fig.1, an additional weight \( d_i \) is incorporated when forming an error signal \( e_k \) to compensate for the non-unitary channel, for which \( e_k \) is short.

We make two further assumptions:

(A2) The channel impulse response \( h_0 \) is non-zero. We will see that this assumption ensures that under ideal adaptation we have \( d_1 = \text{sgn}(h_0) \) for all \( i = 0, 1, \ldots, N \) and thus \( \hat{h}_N = \text{sgn}(h_0) \). Notice that this assumption is not required for the non-unitary channel, which is short.

(A3) The number of DFE weights, \( N+1 \), equals the length of the channel impulse response including the cursor.

Remark: We refer to a decision of the form \( \hat{h}_k = \text{sgn}(h_0) h_k \) as correct even though initially it may seem nonsensical for the case \( h_0 < 0 \). The reader should be convinced that they only satisfy this hypothesis that the last \( N \) decisions are correct imply the next and all future decisions are correct.

We introduce a vector convention to simplify the presentation. It will be necessary to consider vectors belonging to \( \mathbb{F}^N \) and to \( \mathbb{F}^N \). Generally, the DFE output equation (2.1) involves vectors in \( \mathbb{F}^N \), whereas we will see the equations describing adaptation necessarily involve vectors in \( \mathbb{F}^N \) since we are adapting \( N+1 \) weights. Our convention is as follows. Let \( W \) represent a vector in \( \mathbb{F}^N \). Then \( W \in \mathbb{F}^N \) is derived from \( W \in \mathbb{F}^N \), by deleting the first component (i.e., by projection). For example, we define the channel impulse response vector as \( H = [h_0, h_1, \ldots, h_N]^T \) along with its associated vector \( D_k \in \mathbb{F}^N \). The vectors representing past and present data and data estimates are \( \hat{A}_k = [\hat{A}_k, \ldots, \hat{A}_k] \) and \( A_k = [A_k, \ldots, A_k] \). In particular, we will be using \( A_k \) and \( \hat{A}_k \) to represent the past only data and data estimates, respectively.

With these definitions, (2.1) can be written succinctly

\[
\hat{A}_k = \text{sgn}(h_0) A_k + \frac{1}{N} \sum_{l=1}^{N} \hat{A}_l. \tag{2.2}
\]

The form of (2.2) will make the partition of parameter space, treated in section 3, more transparent. Our emphasis in this paper is with the stochastic dynamics of the blind adaptation of DFEs. This means we will be interested in the time-varying properties of the tap weights \( D_k \) described (typically) through the fundamental tap weight update equation

\[
D_{k+1} = D_k + \gamma e_k A_k. \tag{2.3a}
\]

The scalar error \( e_k \) represents the discrepancy between the input and the renormalized output of the slicer described by the \( \text{sgn}(\cdot) \) function, (see Fig.1). The scalar \( \gamma \) in (2.3a) represents a small (time-invariant) adaptation gain. Typically, \( \gamma = 0.01 \). Later we will see that it is the small (but not necessarily infinitesimal) value taken by \( \gamma \) which makes an analysis of adaptation through averaging methods tractable. Note that in equations (2.3a) and (2.3b) we use the decisions \( \{\hat{y}_k\} \) rather than the true data \( \{y_k\} \) when adapting; this is referred to as blind adaptation. In other words, we do not use a training sequence to adapt the DFE tap weights. This results in a considerable complication to the analysis, as analogous results for the case of linear equalization have indicated [5–9].

Equation (2.3a) is in the form of a standard least mean square (LMS) adaptation algorithm [15] (with input \( \{\hat{y}_k\} \) rather than the more usual \( \{y_k\} \)). Equally well we may consider algorithms based on normalized LMS (NLMS) or Recursive Least Squares (RLS) or variations thereof. However, for brevity, we will consider only (2.3a) in detail in this paper.

A major problem we consider in this paper is the interplay between the output equation (2.2) and the adaptive tap weight update algorithm (2.3a). In particular, we will be interested in deriving an averaged equation describing the mean trajectory of the DFE tap weights \( D_k \) in \( \mathbb{F}^N \) (the mean is in the sense of the ensemble of possible input binary sequences). (Typically, the variation about this mean can be described by a diffusion approximation [8]). We also highlight the role played by the channel parameters \( H \) in determining whether or not blind adaptation will lead to difficulties in convergence.

We note that since the subject of our investigations is blind adaptation we need to incorporate error propagation into the analysis, i.e., we cannot assume past decisions are correct. This is the main feature which makes the analysis both difficult and interesting.

3 THE PARTITION OF PARAMETER SPACE

3.1 The General Partition

In this section we quantify the effect of the quantization property of the \( \text{sgn}(\cdot) \) function on the behavior of the DFE. We begin by studying the non-adiative DFE equation (2.2).

Define two concatenated vectors: \( V \in \mathbb{F}^{N+1} : = \hat{D} + \gamma L^T \in \mathbb{F}^{N+1} \) representing a general parameter \( V \)-space, and

\[
X_k = [\hat{A}_k^T, \hat{A}_k^T]^T \in \mathbb{F}^N, \tag{3.1}
\]

which plays the role of a state vector. With the above concatenated vector definitions we can rewrite (2.2) as

\[
\hat{A}_k = \text{sgn}(h_0) A_k + V^T X_k, \tag{3.2}
\]

noting that for a fixed \( X_k \), i.e., a given state, \( V^T X_k = \theta_0 \) is the equation for a hyperplane in \( V \)-space.

Consider now, for sake of argument, \( \hat{A}_k = -1 \) for some fixed state \( X_k \). Then for all \( V \in \mathbb{F}^{N+1} \) satisfying

\[
(V \in \mathbb{F}^{N+1} \mid V^T X_k = -\theta_0), \tag{3.3}
\]

we find the argument of the \( \text{sgn}(\cdot) \) function in (3.2) is zero. Hence, the hyperplane in (3.3) partitions \( V \)-space such that on one side of this hyperplane (the side containing the origin) \( \hat{A}_k = -\text{sgn}(h_0) \), i.e., a correct decision, and on the other side \( \hat{A}_k = +\text{sgn}(h_0) \), i.e., an incorrect decision. Similarly, with \( \hat{A}_k = +1 \) we get a hyperplane parallel to the first in the origin, which can be written

\[
(V \in \mathbb{F}^{N+1} \mid V^T X_k = +\theta_0). \tag{3.4}
\]

Hence for a given state \( X_k \) (or equivalently \( -X_k \)) it is natural to consider the parameter space partitioned into three regions by the two parallel hyperplanes (3.3) and (3.4).

We can repeat this analysis for all \( 4^N \) possible values of the state \( X_k \). There are then \( 4^N \) pairs of hyperplanes subdividing parameter space. However, the total number of distinct hyperplanes taking into account redundancy is \( 4^N \). This gives us our desired complete partition of the parameter space. The (smallest) regions thus produced are defined by the intersection of a number of half-spaces, and therefore form convex polytopes. We conclude: For any two points interior to one such polytope we could not distinguish between the corresponding DFE systems on the basis of
FIGURE 2: The 64 Polytopes in $D_k$-space for the N=2 Channel $h_0=1$, $H=[4\ 3]^T$.

observations at $a_k$. This has wide ranging implications, e.g., during adaptation, only when the trajectory of the tap weights ($D_k$) crosses the boundary of a polytope will we observe a change in the stochastic error dynamics of the DFE and hence the behaviour of adaptation.

Remark: The polytopes above may be thought of as lower dimensional projections of polytopes existing in the parameter space of vectors of the form $[H^1;D_k]^T \in \mathbb{R}^{N+2}$, called W-space, or V-space with $h_0$ and $a_k$ augmented.

3.2 Post-Adaptation Error Propagation -- Partition

The idea of tessellation of parameter space into polytopes was brought up in [1] which investigated the error propagation mechanism of DFEs whose tap weights had converged (other authors have recognised this discrete property of the parameter space, e.g., [16]). Such polytopes are merely the lower dimensional projections of W-space polytopes defined by imposing $D_k = H$, deleting $a_k$ and fixing $h_0 > 0$ (the latter to be compatible with [1], i.e., more generally we would fix $D_k = \text{sgn}(h_0) H$ with $h_0 \neq 0$). Indeed, the derivation of the equations of the hyperplanes defining the polytopes in the previous subsection is a generalization of the derivation found in [1].

3.3 Adaptation to a Time-Invariant Channel -- Partition

The channel parameters $H$ will be fixed for a time-invariant channel. Hence, we have only to consider the space of N+1 feedback tap weights, $D_k$-space (which can be regarded as a function of $H$) when dealing with adaptation to a time-invariant channel. A typical hyperplane that subdivides $D_k$-space (or equivalently $D_k$-space) into polytopes is given by

$$D_k e \in \mathbb{R}^{N+1} \text{ or } D_k e \in \mathbb{R}^{N} \quad | \quad D_k e = A_k h_0 + A_k T h_1,$$

for a given state $X_k$ (remember $X_k \in [A_k^T;A_k^{-1}^T]$) and fixed data $a_k$. The derivation of (3.5) is entirely analogous.

3.4 Example of Polytopes in $D_k$-space for N=2

Let $N = 2$, $h_0 = 1$, $h_2 = 3$. The $a_2$-values are given by $d_1 e = d_2 e = d_3 e = 0$ for $e \in \{0,1,2,3,6,18\}$ according to (3.5). These lines are depicted in Fig.2. Notice that due to the degeneracy in this example we really only have only 14 distinct lines rather than the usual full quota of 16. We will return later to consider this example in more detail.

4. FINITE STATE MARKOV PROCESSES

4.1 Stochastic Modeling of the Non-Adaptive DFE

We need a mechanism whereby we can quantify the effects of a stochastic input $a_k$ on the behaviour of the DFE. Our quest for a suitable description is eased by making a realistic whiteness assumption on the input $a_k$.

(a.4) The input binary sequence $a_k$ is white.

(Note that (A.4) implies $E[a_k a_k^T] = I$.) We also assume a uniform distribution, viz.,

$$E[a_k] = \mathbb{P}(a_k = +1) = \mathbb{P}(a_k = -1) = \frac{1}{2}.$$  

(A.5) The data $a_k$ takes binary values with equal probability.

Then it is a standard result that a finite state Markov process (FSMP) can be developed to describe the stochastic relationship between the decisions $a_k$ and the input $a_k$ [14, 17-19]. Given that the channel and DFE parameters are fixed, however, the FSMP developed in [1] is sufficiently general (capable of handling the case $D_k = \text{sgn}(h_0) H$) to meet our needs for this paper. We will see later that the FSMP describing the non-adaptive stochastic dynamics varies according to the channel and DFE parameters but only in a discrete fashion.

In [1] it was shown that the 4$^N$ values taken on by the atomic state vector $X_k$ (3.1) form Markov states in a FSMP. This set of 4$^N$ Markov states is denoted $\Omega$. This FSMP can simply be viewed as imposing a probabilistic structure onto a state transition diagram which describes all possible transitions between atomic states. Summarizing the development in [1] (see also [19]): Given a "binary" ordering of atomic states indexed by $i \equiv \{0,...,2^{2N-1} - 1\}$, where $i \in \{0,1,2,...,m \leq 4^N\}$, we associate a probability transition matrix $P$ and time-varying (non-stationary) probability density vector $v_k$ such that $v_{k+1} = P v_k$. Typically there exists a unique stationary or steady state distribution $v_\infty \equiv \mathbb{P} v_k$. (Unfortunately, there exist $P$ with non-unique $v_\infty$.) However, it appears that this only occurs for collections of polytopes which adaptation should sensibly avoid, see section 6.3. For our purposes we wish simply to indicate how certain important statistical quantities can be calculated using this FSMP, assuming $v_\infty$ is unique.

4.2 Error Probabilities and the Covariance Matrices

To make calculations tractable, we make the following assumption:

(a.6) All FSMPs considered are assumed stationary.

For definiteness we show how to calculate the various stochastic quantities of interest by considering the case $N = 2$. (Higher order cases are handled similarly, albeit with increased computational burden.) With $N = 2$ there are 16 atomic states. In the binary ordering of the last subsection, atomic state $X_k = [-1 \ 1 \ -1 \ -1]^T$ codes as $i = 1$. Recalling the definition of the components in $X_k$ (3.1), we can compute the error probability $P_{e \equiv \mathbb{P}(a_k \neq \text{sgn}(h_0) A_k)}$ under stationarity by first recording the indices corresponding to either (i) those atomic states whose first component does not equal $\text{sgn}(h_0)$ times the third component, i.e., an error occurred one time step ago, or alternatively (ii) those atomic states whose second component does not equal $\text{sgn}(h_0)$ times the fourth component, i.e., an error occurred two time steps ago. Then steady state $P_{e \equiv \mathbb{P}(a_k \neq \text{sgn}(h_0) A_k)}$ can be computed by adding the components of $\gamma_k$ whose indices were recorded. For example, if $\text{sgn}(h_0) > 0$, then in each case we get

$$P_{e \equiv \mathbb{P}(a_k \neq \text{sgn}(h_0) A_k)} = \gamma_{+1} + \gamma_{+2} + \gamma_{+4} + \gamma_{-1} + \gamma_{-2} + \gamma_{-4}.$$  

The second major class of statistical quantities of interest to us is covariance. In particular we will be interested in Toeplitz matrices of the form $R \equiv E[A_k A_k^T]$ and $C \equiv E[A_k A_k]$, which have scalar components of the form $E[A_m a_k]$, $E[A_k a_k]$, respectively. Then using a simple extension of the argument of the $C_k$ calculation we have, e.g., $E[A_k a_k] = \gamma_{+1} + \gamma_{+2} + \gamma_{+4} + \gamma_{-1} + \gamma_{-2} + \gamma_{-4} = \gamma_{+1} + \gamma_{-1}$. The details are not important. However what is important is that the $R$ and $C$ matrices are determinable from $P$ and $v_\infty$, i.e., the FSMP. We will see in section 5.2 that $R$ and $C$ are crucial parameters in the description of the adaptive stochastic dynamics of DFEs.

5. EQUILIBRIA AND AVERAGING ANALYSIS

5.1 Wiener-Hopf Solution

Our first task is to determine the locations of the
attraction points (equilibrium points) in $D_k$-space for the weights when the algorithm seeks to minimize a least squares criteria as in the LMS algorithm (2.3a).

Consider Fig.3 which simply redraws a portion of Fig.1. (The definitions of the symbols in Fig.3 are identical to those in Fig.1.) This figure suggests that $A_k$ can be interpreted as an input sequence and that $y_k$ can be interpreted as a desired response. As is well known, the objective of the LMS algorithm (2.3a) is to minimize the mean square error defined by $E(D_k) = E(e_k^2(D_k))$ which is a quadratic (hence uni-modal) surface. The tap weight setting, $Deq$, which minimizes the mean square error is the classical Wiener–Hopf formula (15) and is given by

$$ Deq = E[A_k A_k^T]^{-1} E[A_k y_k] $$

where we have defined: (i) the covariance of $A_k \equiv E[A_k A_k^T]$ (non-negative definite), and (ii) the cross-covariance vector $Q \equiv E[A_k y_k]$. Note, from this point on we consider only the case when $R$ is non-singular. We will see that this restriction is justifiable when we come to section 6.3.

In reality, $y_k$ is not a user supplied sequence but simply the channel output which in our case is identified with $y_k = A_k^T H$. Hence (5.1a) may be expanded to

$$ Deq = E[A_k A_k^T]^{-1} E[A_k y_k] $$

This formula for $Deq$ (equilibrium), which gives the minimum mean square error is the classical Wiener–Hopf formula (15) and is given by

$$ Deq = E[A_k A_k^T]^{-1} E[A_k y_k] $$

We will demonstrate that the blind LMS algorithm (2.3a) can be (but is not necessarily) attracted to undesirable regions of $D_k$-space where the channel is not correctly equalized and the error rates are unacceptably high.

The parameters which determine the dynamics of the mean equation (5.4b) are the covariance matrices $R$ and $C$. We also met these matrices earlier in section 4 and we showed they could be evaluated with the assistance of FSMPs. Now, recall our one-to-one correspondence between the polytopes and FSMPs. Hence, in $D_k$-space the matrices $R$ and $C$ will be constant only whilst the tap setting $D_k$ remains inside any one $D_k$-space polytope (assuming steady state). Therefore, we have shown (albeit informally)

**Proposition 1**: In steady state, the matrices $R$ and $C$ are piecewise constant functions of $D_k$, where the “pieces” are precisely the $D_k$-space polytopes defined by the hyperplanes (3.5).

**Remarks**: (i) To emphasize $R$, $C$ and $Q = CH$ are discrete functions of the polytope $P$ for which $D_k \in P$, we write $R(P)$, $C(P)$ and $Q(P)$.

(ii) Within each polytope the mean trajectory (5.4b) is determined by a constant coefficient, linear, deterministic difference equation. Hence, over the whole $D_k$-space, the averaged trajectory describing the complete adaptation is determined by a piecewise constant coefficient, linear, deterministic difference equation (see the example in section 6.2).

The next property is an embellishment on proposition 1.

**Proposition 2**: The mean square error surface $E(D_k) \equiv E(e_k^2(D_k))$ is a piecewise (polytope-wise) quadratic function of $D_k$ given by

$$ E(D_k + P) = E[y_k^2] + 2Q(P)D_k + D_k R(P)D_k $$

**Proof**: This is a trivial modification of a standard result in adaptive least squares, i.e., (6.1a) is [15, eqn (2.31)]. To obtain (6.1b), we substitute both $y_k = A_k^T H$ and $Q(P) = C(P)H$ into (6.1a), and invoke the whiteness of the input (A.4). Of course, we emphasize that for blind adaptation of DFFs, $C(P)$ and $R(P)$ are constant only within a polytope $P$, and generally these matrices vary from polytope to polytope. Therefore, the quadratic surface is different for different polytopes, i.e., piecewise quadratic.

**Remark**: The local minima of the mean square error $E(Deq(P))$ associated with a polytope $P$, in terms of our fixed $H$, can be written $E(Deq(P)) = H^T [C(P)R(P)^{-1}C(P)] H$. However, if $D_k(P)$ lies outside $P$ then this minimum is not attainable. We will see in section 6.3 that the global minimum mean square error is zero at $D_k(P) = \text{sgn}(h_1) H$ for the polytope which contains $\text{sgn}(h_1) H$, and is always attainable in the sense described in the following paragraph.

With each polytope $P$ we have associated an equilibrium $Deq(P) = R(P)^{-1}C(P)H$. To recapitulate, it is natural to
classify two types of equilibria.

Definition: $D_{eq}(P)$ is attainable if $D_{eq}(P) \in P$, otherwise it is unattainable.

If $D_{eq}(P)$ is attainable then $D_{eq}(P)$ will tend to move towards and settle down around it, (whenever $D_{eq}(P)$ is). Otherwise, if $D_{eq}(P)$ is unattainable ($D_{eq}(P) \notin P$) will tend to move towards the boundary $\partial P$ nearest to $D_{eq}(P)$ and thus head onto the adjacent polytope. The example in the next subsection best illustrates these ideas.

6.2 Example of Adaptation and the Averaged Trajectory

As in section 3.4 we choose the example $\nu = 1$, $\alpha = 4$ and $\nu = 3$. In Fig.4a we have plotted a large number of averaged trajectories according to (3.4b) with $\gamma = 0.01$, noting that the R and C matrices are dependent on the polytopes (pictured in Fig.2). Note Fig.4a is the projection of $D_{eq}(P)$ space (2D) of $D_{eq}(P)$ space (3D), therefore some of the averaged trajectories only appear to cross. The starting $d$-component for all trajectories was arbitrarily selected at zero. Naturally the evolution of $d$ during adaptation cannot be discerned in such a figure.

Figure 4b shows the precise sense in which to interpret Fig.4a. It shows an insert of Fig.4a with a single (bold) averaged trajectory and four realizations (i.e., simulations according to (5.3)) which appear to cluster about the averaged trajectory.

Note for this example there are only three attainable equilibria at $[1 \ 4 \ 3]^T$, $[4 \ 3 \ 0]^T$ and $[4.667 \ 5.333 \ 3.667]^T$. In Fig.4a the 2D projections of these equilibria are given by $[4 \ 3]^T$, $[3 \ 0]^T$ and $[5.333 \ 3.667]^T$ and these appear as (local) attraction points for the mean trajectories.

6.3 Approximate and Exact Attainable Equilibria

The equilibria predicted by the theory developed are based only on mean behaviour (5.4b). However we note the driving term of the adaptation equation (2.3a) is given by $e_k (2.3b)$. Error $e_k$ in turn is only identically zero (for all k) when $D_k = D_{eq}$, $\equiv \text{sgn}(h,P)$ (the proof is straightforward and omitted). So we say that only $D_{eq}$ is an exact equilibrium in the sense that $e_k(D_{eq}) = 0$. All other attainable equilibria are termed approximate equilibria and the physical manifestation of an attainable equilibrium with $e_k(D_{eq}) \neq 0$ is that the sequence $(D_k)$ is observed to juggle about (but generally stay in the vicinity of) $D_{eq}(P)$.

The questions arise: (i) regarded as a small noise can the variations in $e_k$ (estimating the gradient) drive $(D_k)$ away from $D_{eq}(P)$ such that $(D_k)$ hits $\partial P$ (the boundary of $P$); and (ii) what is the expected time to do so? In the stochastic process literature [11,12] this is known as an exit problem. These are crucial questions because if the DFE hangs at an equilibrium $D_{eq}(P)$ corresponding to high error rates, e.g., $D_{eq} = [4.667 \ 5.333 \ 3.667]^T$ in Fig.4a, then an unacceptably high exit time has serious practical consequences. These questions are the subject of current research.

6.4 Delay-type Equilibria and Their Attainability

In this section we consider those polytopes in $D_k$-space which yield a decision sequence which is a delay of the input (with a possible sign change) under steady state. We derive necessary (and conjecture sufficient) conditions for attainability of the attraction points of these polytopes for a broad class of adaptation algorithms in terms of the channel parameter.

Let $\delta \neq \text{sgn}(h,P)$, then rewriting (2.1) we have,

$$
\delta_k = \text{sgn}(h_i, a_k, \delta) + U_k(\delta) + V_k(\delta),
$$

where

$$
U_k(\delta) = \sum_{i=1}^{N} d_i (a_k - i - \sigma(\delta) a_{k+6,i-1}) - \frac{\sigma(\delta)}{1+i}\sum_{i=1}^{N} d_i (a_k - i) + \sum_{i=1}^{N} d_i (3.2b)
$$

$$
V_k(\delta) = \sum_{i=1}^{N} \delta_i (a_k - i - \sigma(\delta) a_{k+6,i-1}) - \frac{\sigma(\delta)}{1+i}\sum_{i=1}^{N} d_i (3.2b)
$$

and

$$
V_{max}(\delta) = \sum_{i=1}^{N} \delta_i (a_k - i - \sigma(\delta) a_{k+6,i-1}) - \frac{\sigma(\delta)}{1+i}\sum_{i=1}^{N} d_i (3.2b)
$$

$$
V_{max}(\delta) = \sum_{i=1}^{N} \delta_i (a_k - i - \sigma(\delta) a_{k+6,i-1}) - \frac{\sigma(\delta)}{1+i}\sum_{i=1}^{N} d_i (3.2b)
$$

The reason for the curious decomposition given by (6.2a) will become clearer later. We will see that $\delta$ corresponds to a nominal time delay and of $\delta \neq \text{sgn}(h,P)$, corresponds to an associated sign change of the channel/DFE combination.

We define two proper subsets of the set of atomic Markov states $\Omega$ (section 4.2) parametrized by $0 < \delta < N$. (The case $\delta = N$ needs to be treated separately but fortunately is easily disposed of.) Define,

$$
\omega_0(\delta) = \{X_k \in \partial a_k - i - \sigma(\delta) a_{k+6,i-1}, 1 \leq i \leq 6, N-\delta
$$

$$
\omega_{-\delta}(\delta) = \{X_k \in \partial a_k - i - \sigma(\delta) a_{k+6,i-1}, 1 \leq i \leq 6, N-\delta
$$

both of which consist of collections of $2N+\delta$ atomic states where (precisely) the $N-\delta$ most recent decisions are of the form $\partial a_0 = \omega(\delta)$ and $\delta$, respectively.

Remark: Note the definitions of $\omega(\delta)$ simply express that the member atomic states vectors have their $N-\delta$ component equal to $\omega(\delta)$ times the $\delta$-th component for $i = 1, ..., N-\delta$, and thus these subsets are in effect independent of $k$ (as the notation suggests).

The notion of a closed subset of $\Omega$ will considerably simplify development.
Definition: A subset of $\Omega$ is closed if transitions from any one atomic state in the subset is only to another within the subset.

The following statements are equivalent: (i) Suppose $X_k \in \omega_0$, then all future $(m \geq k)$ decisions are of the form $\tilde{a}_m = \tau(\delta) a_{m-\delta}$. (ii) $\omega_0$ is closed. Hence to investigate channel/DFE combinations yielding simple time delay behaviour we need only to determine when $\omega_0$ are closed and reachable from within $\Omega$. Our first proposition narrows our investigations by showing a DFE can never behave consistently in the form $\tilde{a}_m = -\tau(\delta) a_{m-\delta}$ and in a steady state stochastic environment, and it also gives necessary and sufficient conditions for $\omega_+(\delta)$ closure.

Proposition 3: (a) $\omega_+(\delta)$ is closed if and only if
\[ |h_k| > V_{\text{max}}(\delta). \]  
(b) $\omega_-(\delta)$ is never a closed subset.

Proof: Suppose $\omega_+(\delta)$ is closed and that the system has been in $\omega_+(\delta)$ for some time. Then $\tilde{a}_{k-i} = \tau(\delta) a_{k-i-1}$ for $i = 1, 2, ..., N$. Substituting into (6.2b) and (6.2c) we obtain
\[ |V_k| = 2 \sigma(\delta) \sum_{i=1}^{N} a_{k-i-1} \left| h_k \right| < 0, \quad (6.5a) \]
and
\[ \tilde{a}_{k-1} N \]
\[ V_k^{\pm}(\delta) = \sum_{i=1}^{N} \left| h_k \right| T \delta i \sigma(\delta) a_{k-i-1} \text{sgn}(d_{k-i-1}) \left| d_{k-i-1} \right| > 0, \quad (6.5b) \]

Consider first $\omega_-(\delta)$. If $\left| h_k \right| > V_{\text{max}}(\delta)$ with $V_k^{\pm}(\delta) = 0$ then this implies $\left| h_k \right| > V_{\text{max}}(\delta)$ hence $\omega_-(\delta)$ is closed and the necessary and sufficient condition (6.4) is also necessary. Now consider $\omega_+(\delta)$. In this case we have $\tilde{a}_{k-i} = \text{sgn}(d_{k-i-1}) + V_k^{\pm}(\delta) + V_{\text{max}}(\delta)$ where by (6.5a,b) $V_k^{\pm}(\delta) + V_{\text{max}}(\delta)$ is independent of $a_{k-i-1}$. Hence $\text{Pr} = (0, \text{sgn}(d_{k-i-1}))$ is a contradiction. 

Contriving an input sequence which visits all atomic states in $\omega_-(\delta)$ when the initial state is in $\omega_+(\delta)$ is straightforward. This shows no proper subset of $\omega_+(\delta)$ is closed assuming $\omega_+(\delta)$ itself is closed. We formulate this as:

Proposition 4: $\omega_+(\delta)$ is indecomposable.

The inequality (6.4) can only hold for at most one value of $l$. To prove this one assumes at least two inequalities of the form (6.4) are simultaneously satisfied (say for $s_1$ and $s_2$) then an application of the triangle inequality establishes a contradiction. The details of the proof are omitted. We state this result as Proposition 5.

Proposition 5: $\omega_+(\delta)$ is closed for at most one $\delta$.

Proposition 6: Let $h_k$, $0 \leq \delta < N$ satisfy (6.4). Then the following alternative conditions are sufficient to guarantee that there exists an input sequence such that $N^{\delta}$ consecutive $\tilde{a}_m = \tau(\delta) a_{m-\delta}$ decisions are made:
(i) $\delta = 0$, $N = 2$, $N = 1$, $N$ (and thus cases $N = 1, 2, 3$)
(ii) $\text{sgn}(d_{k}) \neq \text{sgn}(d_{k-\delta}) = 1$.

Conjecture 7: Let $h_k$, $0 \leq \delta < N$ satisfy (6.4). Then there exists an input sequence such that $N^{\delta}$ consecutive decisions are made of the form $\tilde{a}_m = \tau(\delta) a_{m-\delta}$.

Remarks: (i) With (6.4) satisfied, and the hypothesis of Proposition 6 fulfilled, $\omega_+(\delta)$ is closed, indecomposable and reachable, so that $\text{Pr}(X_k \in \omega_+(\delta)) = 1$ as $\delta = 1$. Hence under stationarity the channel/DFE combination produces decisions of the form $\tilde{a}_m = \tau(\delta) a_{m-\delta}$ if and only if $|h_k| > V_{\text{max}}(\delta)$ with $\omega_+(\delta)$ reachable. Note that the output is clearly white under such conditions.

(ii) Given a time-invariant channel $H$, define the following region of $D_k$-space (which is just another way of writing (6.3)),
\[ \epsilon(\delta) \geq (D_k \epsilon)^{N+1} \rho(\delta) \sum_{i=1}^{N} |h_i| - \tau(\delta) a_{i-\delta} > 0, \quad (6.6a) \]
where
\[ \rho(\delta) = \frac{|h_i|}{\delta - 1} \sum_{i=0}^{N} |h_i| 0 \leq \delta < N. \]

Then $\tilde{a}_m = \tau(\delta) a_{m-\delta}$ under steady state conditions only if $D_k \in \overline{\Omega(\delta)}$ and sometimes if (according to the reachability of $\omega_+(\delta)$). Note this region may also be written $\rho(\delta) > |D_k - \epsilon|^{N+1} \langle \langle \Omega(\delta) \rangle \rangle$, where $\langle \langle \cdot \rangle \rangle$ denotes an $n$-dimensional hypercube.

Remarks: (i) $|H|$, is the peak excursion of the channel output when driven by a white binary input. Hence we can estimate $|H|$, by channel output measurements and thus impose during adaptation that $\tilde{a}_m$ does not leave $N(\delta) = \epsilon^2$, etc.

6.5 White Equilibria

As we have earlier commented, when $\tilde{a}_k = \tau(\delta) a_{k-\delta}$, the $\epsilon(\delta)$ process is white. Let us term any equilibrium with the $\epsilon(\delta)$ process white as a white equilibrium. In this subsection, we shall present further results on this class.

We now give two closely related propositions which imply that adaptation should be restricted to a well defined region of $D_k$-space.

Proposition 9: Suppose $\tilde{a}_k$ is white, (under steady state). Then
\[ \left( h_i \right) \text{sgn}(d_{k}) > 1 = D_k \left( \text{sgn}(d_{k}) \right), \quad (6.7) \]

Proof: If $\tilde{a}_k$ is white then the subsequence $\tilde{a}_k = -1(\delta) a_{k-\delta}$, $\delta = 1, 2, ..., N$ occurs with non-zero probability (for some input sequence). In (2.1) this implies $\tau(\delta) = -1 = \text{sgn}(d_{k}^2 + D_k^2)$. (6.7) follows.

Remarks: (i) $|H|$, is the peak excursion of the channel output when driven by a white binary input. Hence we can estimate $|H|$, by channel output measurements and thus impose during adaptation that $\tilde{a}_m$ does not leave $N(\delta)$. 


Clearly, by the definition of adaptation algorithms should constrain almost surely. Then in the two subsequences mite can contradict because otherwise we would have evidence that the same region implies the stationary atomic selections given certain hypotheses, testing is far easier, as indicated (ii) Correlation: Given (i) Zero mean: \( \text{E}(x_k) = 0 \) for all \( k \geq 0 \). (iii) Sign symmetry: \( \text{Pr}(x_k = x_k) = \text{Pr}(x_k = -x_k) = \frac{1}{2} \). (iv) No parity bias: the probability that for any sample of size \( 2m \) of \( x_k \) having an even number of \( +1 \)'s equals the probability that the same sample has an odd number of \( -1 \)'s, then the \( x_k \) process is white.

As already noted, delay equilibria are white. We suspect the reverse implication is also true, and formalize the statement as:

**Conjecture 12**: If \( x_k \) is white then \( x_k = \pm x_k \) for precisely one \( \pm \) and \( \neq N \). If the conjecture held we would have a way of statistically testing the output of a DFE to prove it was inverting the channel with suitable delay. Note also for a very large class of practical channels \( V_{\text{med}}(x) \geq 10k_1 \) for all \( x \geq 0 \). Thus the only possibility in this case is to have \( s = 0 \) which is a highly desirable situation.

**7 CONCLUSIONS**

In this paper, we have described some analytical tools which should be of assistance in dealing with the fundamental question: for what classes of channel impulse responses if DFE use a practical proposition? We are not yet in a position to answer this question, and we have indicated several conjectures bearing on the use of these tools. Let us note here two directions in which considerably more work may need to be done to move in the direction of a definitive answer. First, the techniques of large deviation analysis [10,11] need to be brought to bear to consider how long tapers are likely to remain in the neighbourhood of an incorrect or inappropriate minimum. Second, it would be highly desirable to find some systematic way of approximating a Markov process with a much smaller number of states (and we note here the work of [21]), or by a Markov process with continuous state space for which analytic calculation could be more easily executed (there is really no evidence to suggest this would be possible for the DFE problem, but the possibility cannot be excluded a priori).

**REFERENCES:**


[16] P.V. Kaballa (private communication).


