PART II

8. Structure-Related Issues In Adaptive Filtering

The input-output characteristic of linear systems is classically described by a ratio of polynomials in shift operator notation. However, there exists an infinite number of "equivalent" descriptions (also referred to as realizations, representations, or parameterizations) with the same external (input-output) behavior, but different internal behaviors. These equivalent descriptions are most commonly known as the state-space representations. Different state-space representations have different sensitivity measures and different numerical properties under finite-word-length conditions [9]. Furthermore, the state-space adaptive filters are based on matrix operations, which can be easily implemented by systolic arrays [32]. This motivates us to explore the possibility of using different structures in adaptive filtering and to investigate whether the structural properties can be exploited to improve adaptive filter performance.

8.1 Motivation

An N^{th} order linear system is described in state-space form by the following equations:

$$\mathbf{x}_{n+1} = \mathbf{A}_n \mathbf{x}_n + \mathbf{B}_n \mathbf{u}_n \tag{8.1a}$$

$$\mathbf{y}_n = \mathbf{C}_n \mathbf{x}_n + \mathbf{D}_n \mathbf{u}_n \tag{8.1b}$$

where $\mathbf{x}_n$ is a vector of N states, $\mathbf{u}_n$ is a $q \times 1$ vector of inputs, $\mathbf{y}_n$ is a $p \times 1$ vector of outputs, $\mathbf{A}_n$ is a matrix of dimension $N \times N$, $\mathbf{B}_n$ is a matrix of dimension $N \times p$, $\mathbf{C}_n$ is a matrix of dimension $q \times N$, and $\mathbf{A}_n$ is a matrix of dimension $q \times p$. In this dissertation, we focus mainly on time-invariant single-input-single-output (SISO) systems. Hence, the dependence of the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ on n (as indicated in (8.1) with a suffix) is dropped and we assume $p = q = 1$. We also assume that the system order N is known *a priori*. The state-space representation $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ is not unique. For any non-singular matrix $\mathbf{T}$ of dimension $N \times N$, the state-space representation $\{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \mathbf{T}^{-1}\mathbf{B}, \mathbf{C}\mathbf{T}, \mathbf{D}\}$ has the same input-output behavior as (8.1).

Let $\boldsymbol{\theta}$ be a finite dimensional parameter vector that uniquely defines the state-space matrices. That is, there exists a one-to-one mapping between $\boldsymbol{\theta}$ and $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$. Identifying the class of

uniquely identifiable state-space parameterizations is also an important structure related issue in adaptive filtering [35]. We rewrite the state-space equations (8.1) as follows.

$$\mathbf{x}_{n+1} = \mathbf{A}(\boldsymbol{\theta})\mathbf{x}_n + \mathbf{B}(\boldsymbol{\theta})\mathbf{u}_n \tag{8.2a}$$

$$\mathbf{y}_n = \mathbf{C}(\boldsymbol{\theta})\mathbf{x}_n + \mathbf{D}(\boldsymbol{\theta})\mathbf{u}_n \tag{8.2b}$$

The dependence of the state-space matrices on the parameterization is explicitly shown in (8.2).

Assume that L input-output data pairs are available. Let us say that a least squares method is used to estimate $\boldsymbol{\theta}$. The estimate $\hat{\boldsymbol{\theta}}_L$ provided by the least squares method is the value of $\boldsymbol{\theta}$ that minimizes the average squared prediction error

$$V_L(\boldsymbol{\theta}) = \frac{1}{L}\sum_{n=1}^{L}(e_n(\boldsymbol{\theta}))^2 \tag{8.3}$$

where $e_n(\boldsymbol{\theta})$ is the prediction error given by

$$e_n(\boldsymbol{\theta}) = y_n - \hat{y}_{n/\boldsymbol{\theta}} \tag{8.4}$$

and $\hat{y}_{n/\boldsymbol{\theta}}$ is the predicted output according to

$$\hat{\mathbf{x}}_{n+1/\boldsymbol{\theta}} = \mathbf{A}(\boldsymbol{\theta})\hat{\mathbf{x}}_{n/\boldsymbol{\theta}} + \mathbf{B}(\boldsymbol{\theta})\mathbf{u}_n \tag{8.5a}$$

$$\hat{\mathbf{y}}_{n/\boldsymbol{\theta}} = \mathbf{C}(\boldsymbol{\theta})\hat{\mathbf{x}}_n + \mathbf{D}(\boldsymbol{\theta})\mathbf{u}_n \tag{8.5b}$$

From an abstract viewpoint, the parameterization we decide to use has no special significance, other than serving as a vehicle to arrive at a good input-output description of the system. However, from a practical viewpoint, the parameterization decides how easy it is to *numerically* minimize the cost function shown in (8.3). The answer to this question depends partly on the properties of the Hessian (second derivative matrix)

$$\frac{d^2}{d\boldsymbol{\theta}^2}V_L(\theta) \stackrel{\Delta}{=} V_L''(\theta) \tag{8.6}$$

The conditioning of the Hessian is an important factor in the minimization problem for different reasons. Firstly, the Hessian describes the local character of the minimum. Secondly, the inverse of the Hessian is used in Newton-type minimization algorithms such as RLS. Furthermore, if a simple gradient-type algorithm such as LMS is used to estimate the parameters, the convergence rate of the algorithm also depends on the condition number of the Hessian. In particular, the convergence is very slow if the Hessian is badly conditioned, even if this is locally "on the way

to the minimum." In addition to being well conditioned, it is desirable to have a Hessian that is diagonal or with many zero off-diagonal elements [15]. This is because the negative of the gradient points towards the minimum, if the Hessian is diagonal and well conditioned. On the other hand, if the Hessian has significant off-diagonal elements, the gradient does not point towards the minimum. Furthermore, off-diagonal elements introduce coupling between different coefficients in the parameter vector. Even if one of the coefficients, say the j^{th} coefficient, has reached its optimal value, it would move away from the optimal value if any other coefficient, say the i^{th} coefficient, has not reached its optimal value, unless $\left[V_L''(\boldsymbol{\theta})\right]_{i,j} = 0$.

The properties of the Hessian, as evident from (8.2) - (8.6) depend on $\boldsymbol{\theta}$, the parameterization used. Hence, it is worthwhile to check if there exists an "optimal" structure that can be used to realize an adaptive filter with "optimal" performance.

8.2 Sensitivity

The parameter sensitivity of the structure also has an effect on the performance of adaptive filters [10]. System parameter sensitivity is defined as

$$S = \sum_{i=1}^{NP} \frac{1}{2\pi j} \oint \frac{\partial H(z,\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial H(z^{-1},\boldsymbol{\theta})}{\partial \theta_i} \frac{dz}{z} \tag{8.7}$$

where $H(z,\boldsymbol{\theta})$ is the (stable) system transfer function and $NP(=\dim \boldsymbol{\theta})$ is the number of system parameters. The transfer function of the state-space system shown in (8.2) is given by

$$H(z,\boldsymbol{\theta}) = C(\boldsymbol{\theta})[zI - A(\boldsymbol{\theta})]^{-1} B(\boldsymbol{\theta}) + D(\boldsymbol{\theta}) \tag{8.8}$$

The sensitivity measure shown in (8.7) is a measure of the frequency response deviation caused by small changes in the parameters. It is evident from (8.7) and (8.8) that sensitivity also depends on the parameterization (structure) chosen. Hence, one would expect a structure with high parameter sensitivity to exhibit fast convergence. Similarly, a structure with less sensitivity is expected to have good noise rejection characteristics (robust in the presence of noise), since the wrong parameter estimates, due to the misadjustment caused by noise, will describe a model that is still close to the true system. This conjecture has been verified through simulations by DeBrunner [10]. Thus, the sensitivity viewpoint also encourages investigating the effect of structures on adaptive filter performance.

8.3 Role of the Fisher Information Matrix and the Controllability Grammian

The properties of the Hessian are closely related to the Fisher information matrix and the controllability Grammian. It is worthwhile to study this relationship, since, for example, the properties of controllability Grammians of certain structures such as the lattice structure and the balanced realization are well understood. This section, based mainly on the work done by Van Overbeek and Ljung [33], discusses the relation between the three matrices.

Let us define

$$\bar{\mathbf{J}}(\boldsymbol{\theta}) = EV_L''(\boldsymbol{\theta}) \tag{8.9}$$

where the expectation is over $\{y_n\}$. Assuming that $\{y_n\}$ is stationary, the dependence of $\bar{\mathbf{J}}(\boldsymbol{\theta})$ on L disappears. Define another matrix

$$\mathbf{J}(\theta) = E\psi_{n,\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}}^{-1}\psi_{n,\boldsymbol{\theta}} \tag{8.10a}$$

where

$$\psi_{n,\boldsymbol{\theta}} = \frac{d}{d\boldsymbol{\theta}}\hat{y}_{n,\boldsymbol{\theta}} \tag{8.10b}$$

and

$$\lambda_{\boldsymbol{\theta}} = Ee_{n,\boldsymbol{\theta}}^2 \tag{8.10c}$$

Here also the expectation is over $\{y_n\}$. Taking two successive derivatives on both sides of (8.3) and invoking (8.4), we get

$$V_L''(\boldsymbol{\theta}) = 2\left(\frac{d\hat{y}_{n,\boldsymbol{\theta}}}{d\boldsymbol{\theta}}\right)\left(\frac{d\hat{y}_{n,\boldsymbol{\theta}}}{d\boldsymbol{\theta}}\right)^t - 2e_{n,\boldsymbol{\theta}}\left(\frac{d^2\hat{y}_{n,\boldsymbol{\theta}}}{d\boldsymbol{\theta}^2}\right) \tag{8.11}$$

Suppose there exists a true parameter vector $\boldsymbol{\theta}_0$ such that $\{e_{n,\boldsymbol{\theta}_0}\}$ is a sequence of independent, zero mean random values of variance $\lambda_{\boldsymbol{\theta}_0}$, then from (8.9)-(8.11) we deduce

$$\mathbf{J}(\boldsymbol{\theta}_0) = \lambda_{\boldsymbol{\theta}_0}\bar{\mathbf{J}}(\boldsymbol{\theta}_0) \tag{8.12}$$

Hence, for values of $\boldsymbol{\theta}$ "sufficiently" close to $\boldsymbol{\theta}_0$, the conditioning of the Hessian $\bar{\mathbf{J}}(\boldsymbol{\theta})$ is closely related to the conditioning of $\mathbf{J}(\boldsymbol{\theta})$. If we introduce the assumption that the measurement noise is

Gaussian, $\mathbf{J}(\boldsymbol{\theta}_0)$ is the Fisher information matrix [36]. Hence, the asymptotic covariance of $\hat{\boldsymbol{\theta}}_L$ is proportional to $\mathbf{J}(\boldsymbol{\theta}_0)^{-1}$.

It has been proved that the determinant of $\mathbf{J}(\boldsymbol{\theta})$ is independent of the parameterization used [34]. Since the determinant equals the product of the eigenvalues [17], the smallest eigenvalue of $\mathbf{J}(\boldsymbol{\theta})$ is a measure of how badly $\mathbf{J}(\boldsymbol{\theta})$ is conditioned (equivalently, how bad the parameterization is). Interestingly, the smallest eigenvalue of $\mathbf{J}(\boldsymbol{\theta})$ is related to the smallest eigenvalue of the controllability Grammian $\mathbf{K}(\boldsymbol{\theta})$. The controllability Grammian is also referred to as the state covariance in the signal processing literature, since $\mathbf{K}(\boldsymbol{\theta})$ is the covariance of the states assuming that the input $\mathbf{u}_n$ is white noise of unit variance. The controllability Grammian is the solution to the Lyapunov equation

$$\mathbf{K}(\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{A}^T(\boldsymbol{\theta}) + \mathbf{B}(\boldsymbol{\theta})\mathbf{B}^T(\boldsymbol{\theta}) \tag{8.13}$$

It has been proved that [33]

$$0 \le \underline{C} \le \frac{\lambda_{\min}(\mathbf{J}(\boldsymbol{\theta}))}{\lambda_{\min}(\mathbf{K}(\boldsymbol{\theta}))} \le \overline{C} < \infty, \tag{8.14}$$

where $\lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue of the matrix $\mathbf{A}$ and $\underline{C}$ and $\overline{C}$ are real constants independent of $\boldsymbol{\theta}$, but possibly dependent on $\boldsymbol{\theta}_0$.

8.4 Conclusion

Based on the arguments presented in this chapter, we can conclude that the structure that minimizes the sensitivity measure will have good steady-state properties (misadjustment) and that the structure that minimizes the condition number of the Grammians will have good convergence properties. Interestingly, there is one structure, namely the balanced realization, which has both of these desirable properties. Chapter 9 describes the balanced realization and develops an adaptive IIR filtering algorithm based on this structure.