7.2.2 PARAMETER ESTIMATION VIA PREDICTION ERROR MINIMIZATION

7.2.2.1 Prediction Error Method

The optimal one-step ahead predictor based on the model (7.1) can be written as

$$y(k|k-1) = G(q,\theta)u(k) + \left(I - H^{-1}(q,\theta)\right)\left(y(k) - G(q,\theta)u(k)\right) \qquad (7.19)$$

By comparing (7.1) with (7.19), we see that the prediction error $(y(k) - y(k|k-1))$ is simply $\varepsilon(k)$, assuming that the model is perfect. Note that $\left(I - H^{-1}(q,\theta)\right)$ contains at least one delay since $I - H^{-1}(\infty,\theta) = 0$. Hence, the right hand side does not require $y(k)$ to be known.

Because the primary function of a model in control is to provide a prediction of the future output behavior, it is logical to choose θ such that the prediction error resulting from the model is minimized for the available data record. Let us denote the data record we have as $(\hat{y}(1), \cdots, \hat{y}_N)$. Then, this objective is formulated as

$$\min_{\theta} \sum_{k=1}^{N} \|\hat{e}_{pred}(k,\theta)\|_2^2 \qquad (7.20)$$

where $\hat{e}_{pred}(k,\theta) = \hat{y}(k) - y(k|k-1)$, and $\|\cdot\|_2$ denotes the Euclidean norm. Use of other norms are possible, but the 2-norm is by far the most popular choice. Using (7.19), we can write

$$\hat{e}_{pred}(k,\theta) = H^{-1}(q,\theta)\left(\hat{y}(k) - G(q,\theta)u(k)\right) \qquad (7.21)$$

For certain model structures, the 2-norm minimization of prediction error is

formulated as a linear least-squares problem. For example, for the ARX structure, $G(q,\theta) = \frac{B(q)}{A(q)}$, and $H(q,\theta) = \frac{1}{A(q)}$ and

$$\begin{aligned} \hat{e}_{pred}(k,\theta) &= A(q)\hat{y}(k) - B(q)u(k) \\ &= \hat{y}(k) + a_1\hat{y}(k-1) + \cdots + a_n\hat{y}(k-n) - b_1u(k-1) - \cdots - b_mu(k-m) \end{aligned} \tag{7.22}$$

Since $\hat{e}_{pred}(k,\theta)$ is linear with respect to the unknown parameters, the minimization of
$\sum_{k=1}^{N} \hat{e}^2_{pred}(k,\theta)$ is a linear least squares problem.

Another such example is an FIR model with known disturbance characteristics for which $G(q,\theta) = \sum_{i=1}^{n} h_i q^{-i}$ and $H(q)$ contains no unknown parameters. In this case

$$\hat{e}_{pred}(k,\theta) = \hat{y}_f(k) - h_1u_f(k-1) - \cdots - h_nu_f(k-n) \tag{7.23}$$

where $\hat{y}_f(k) = H^{-1}(q)\hat{y}(k)$ and $u_f(k) = H^{-1}(q)u(k)$. Again, the expression is linear in the unknowns and the prediction error minimization (PEM) is a linear least squares problem. If the noise model was $\frac{1}{1-q^{-1}}H(q)$, then $\hat{y}_f(k)$ and $u_f(k)$ should be redefined as $H^{-1}(q)\Delta\hat{y}(k)$ and $H^{-1}(q)\Delta u(k)$ respectively. The same observation applies to Laguerre or other orthogonal expansion models.

PEM for other model structures such as the ARMAX and Box-Jenkins structures is not a linear least squares problem and pseudo-linear regression is often used for them.

7.2.2.2 Properties of Linear Least Squares Identification

We saw that prediction error minimization for many model structures can be cast as a linear regression problem. The general linear regression problem can be written as

$$\hat{y}(k) = \phi^T(k)\theta + e(k,\theta) \tag{7.24}$$

where $\hat{y}$ is the observed output (or filtered output), ϕ is the regressor vector, θ is the parameter vector to be identified, and e the residual error (that depends on the choice of θ). $\{\cdot\}(k)$ denotes the k_{th} sample. In the least squares identification, θ is found such that the sum of squares of the residuals is minimized, i.e., $\theta_N^{LS} = \arg\left\{\min_\theta \sum_{k=1}^N e^2(k,\theta)\right\}$. We saw in the previous section that 2-norm minimization of prediction error for certain model structures can be cast in this form.

For a data set collected over N sample intervals, (7.24) can be written collectively as the following set of linear equations:

$$\hat{Y}_N = \Phi_N\theta + E_N \tag{7.25}$$

where

$$\Phi_N = \left[\begin{array}{ccc} \phi(1) & \cdots & \phi(N) \end{array}\right]^T \tag{7.26}$$

$$\hat{Y}_N = \left[\begin{array}{ccc} \hat{y}(1) & \cdots & \hat{y}(N) \end{array}\right]^T \tag{7.27}$$

$$E_N = \left[\begin{array}{ccc} e(1) & \cdots & e(N) \end{array}\right]^T \tag{7.28}$$

The least squares solution is

$$\hat{\theta}_N^{LS} = (\Phi_N^T\Phi_N)^{-1}\Phi_N^T Y_N \tag{7.29}$$

Convergence

Let us assume that the underlying system (from which the data are

generated) is represented by the model

$$y(k) = \phi^T(k)\theta_o + \varepsilon(k) \tag{7.30}$$

where θ_o is the true parameter vector (which is assumed to be well defined since we are discussing the convergence here) and $\varepsilon(k)$ is a term due to disturbance, noise, etc.

Some insight can be drawn by rewriting the least squares solution in the following form:

$$\begin{aligned} \hat{\theta}_N^{LS} &= \left[\tfrac{1}{N}\Sigma_{k=1}^N \phi(k)\phi^T(k)\right]^{-1} \tfrac{1}{N}\Sigma_{k=1}^N \phi(k)\left[\phi^T(k)\theta_o + \varepsilon(k)\right] \\ &= \theta_o + \left[\tfrac{1}{N}\Sigma_{k=1}^N \phi(k)\phi^T(k)\right]^{-1} \tfrac{1}{N}\Sigma_{k=1}^N \phi(k)\varepsilon(k) \end{aligned} \tag{7.31}$$

A desirable property of $\hat{\theta}_N^{LS}$ is that under fairly mild assumptions it converges to θ_o as the number of data points becomes large ($N \to \infty$). Note that the term

$$\left[\frac{1}{N}\sum_{k=1}^N \phi(k)\phi^T(k)\right]^{-1} \frac{1}{N}\sum_{k=1}^N \phi(k)\varepsilon(k)$$

represents the error in the parameter estimate. Assume that

$$\lim_{N\to\infty}\left(\frac{1}{N}\sum_{k=1}^N \phi(k)\phi^T(k)\right)$$

exists. This is true if the input is a quasi-stationary signal. In order that

$$\lim_{N\to\infty}\left[\frac{1}{N}\sum_{k=1}^N \phi(k)\phi^T(k)\right]^{-1} \frac{1}{N}\sum_{k=1}^N \phi(k)\varepsilon(k) = 0 \tag{7.32}$$

the following two conditions must be satisfied:

1.
$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^N \phi(k)\varepsilon(k) = 0 \tag{7.33}$$

2.
$$\text{rank}\left\{\lim_{N\to\infty}\left[\frac{1}{N}\sum_{k=1}^{N}\phi(k)\phi^T(k)\right]\right\} = \dim\{\phi\} \tag{7.34}$$

The first condition is satisfied if the regressor vector and the residual sequences are uncorrelated. There are two scenarios under which this condition holds:

- $\varepsilon(k)$ *is a zero-mean white sequence.* Since $\phi(k)$ does not contain $\varepsilon(k)$, $E\{\phi(k)\varepsilon(k)\} = 0$ and $\frac{1}{N}\sum_{k=1}^{N}\phi(k)\varepsilon(k) \to 0$ as $N \to \infty$. In the prediction error minimization, if the model structure is unbiased, $\varepsilon(k)$ is white.
- $\phi(k)$ *and* $\varepsilon(k)$*are independent sequences and one of them is zero-mean.* For instance, in the case of an FIR model (or an orthogonal expansion model), $\phi(k)$ contains inputs only and is therefore independent of $\varepsilon(k)$ whether it is white or nonwhite. This means that the FIR parameters can be made to converge to the true values even if the disturbance transfer function $H(q)$ is not known perfectly (resulting in nonwhite prediction errors), as long as $u_f(k)$ is designed to be zero-mean and independent of $\varepsilon(k)$. The same is not true for an ARX model since $\phi(k)$ contains past outputs that are correlated with a nonwhite $\varepsilon(k)$.

In order for the second condition to be satisfied, $\lim_{N\to\infty}\left[\frac{1}{N}\sum_{k=1}^{N}\phi(k)\phi^T(k)\right]$ must exist and should be nonsingular. The rank condition on the matrix $\lim_{N\to\infty}\left[\frac{1}{N}\sum_{k=1}^{N}\phi(k)\phi^T(k)\right]$ is called the *persistent excitation* condition as it is closely related to the notion of *order of persistent excitation* (of an input signal) that we shall discuss in Section 7.2.2.3.

Statistical Properties

Let us again assume that the underlying system is represented by (7.30). We further assume that $\varepsilon(k)$ is an independent, identically distributed

(i.i.d.) random variable sequence of zero mean and variance r_ε. Then, using (7.31), we can easily see that

$$E\{\hat{\theta}_N^{LS} - \theta_0\} = E\left\{\frac{1}{N}\left(\sum_{k=1}^{N}\phi(k)\phi^T(k)\right)^{-1}\frac{1}{N}\sum_{k=1}^{N}\phi(k)\varepsilon(k)\right\} = 0 \tag{7.35}$$

and

$$\begin{aligned} & E\{(\hat{\theta}_N^{LS} - \theta_o)(\hat{\theta}_N^{LS} - \theta_o)^T\} \\ = & \left(\tfrac{1}{N}\Sigma_{k=1}^{N}\phi(k)\phi^T(k)\right)^{-1}\left(\tfrac{1}{N^2}\Sigma_{k=1}^{N}\phi(k)r_\varepsilon\phi^T(k)\right)\left(\tfrac{1}{N}\Sigma_{k=1}^{N}\phi(k)\phi^T(k)\right)^{-1}]] \\ = & \left(\tfrac{1}{N}\Sigma_{k=1}^{N}\phi(k)\phi^T(k)\right)^{-1}\tfrac{r_\varepsilon}{N} \\ = & r_\varepsilon(\Phi_N^T\Phi_N)^{-1} \end{aligned} \tag{7.36}$$

(7.35) implies that the least squares estimate is "unbiased." (7.36) defines the covariance of the parameter estimate. This information can be used to compute confidence intervals. For instance, when normal distribution is assumed, one can compute an ellipsoid corresponding to a specific confidence level.

7.2.2.3 Persistency of Excitation

In the linear least squares identification, in order for parameters to converge to true values in the presence of noise, we must have

$$\text{rank}\left\{\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\phi(k)\phi^T(k)\right\} = \dim\{\phi\} \tag{7.37}$$

This condition is closely related to the so called *persistency of excitation.* A signal $u(k)$ is said to b *persistently exciting of order* n if the following condition is satsified:

$$\text{rank}\{C_u^n\} = n \tag{7.38}$$

where

$$C_u^n = \lim_{N\to\infty} \frac{1}{N} \Sigma_{k=1}^{N} \left\{ \begin{bmatrix} u(k-1)u(k-1) & u(k-1)u(k-2) & \cdots & u(k-1)u(k-n) \\ u(k-2)u(k-1) & u(k-2)u(k-2) & \cdots & u(k-2)u(k-n) \\ \vdots & \ddots & \ddots & \vdots \\ u(k-n)u(k-1) & u(k-n)u(k-2) & \cdots & u(k-n)u(k-n) \end{bmatrix} \right\} \tag{7.39}$$

The above is equivalent to requiring the power spectrum of $u(k)$ to be nonzero at n or more distinct frequency points between $-\pi$ and π.

Now, suppose $\phi(k)$ consists of past inputs and outputs. A necessary and sufficient condition for (7.37) to hold is that:

$u(k)$ is persistently exciting of order $\dim\{\phi\}$.

This is obvious in the case that $\phi(k)$ is made of n past inputs only (as in FIR models). In this case,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \phi(k)\phi^T(k) = C_u^n \tag{7.40}$$

The condition also holds when $\phi(k)$ contains filtered past inputs $u_f(k-1), \cdots, u_f(k-n)$ (where $u_f(k) = H^{-1}(q)u(k)$). Note that:

$$\Phi_{u_f}(\omega) = \frac{\Phi_u(\omega)}{|H(e^{j\omega})|^2} \tag{7.41}$$

Hence, if $u(k)$ is persistently exciting of order n, so is $u_f(k)$. What is not so obvious (but can be proven) is that the above holds even when $\phi(k)$ contains past outputs.

An important conclusion that we can draw from this is that, in order to assure convergence of parameter estimates to true values, we must design the input signal $u(k)$ to be persistently exciting of order $\dim\{\theta\}$. A pulse is

not persistently exciting of any order since the rank of the matrix C_u^1 for such a signal is zero. A step signal is persistently exciting of order 1. A single step test is inadequate in the presence of significant disturbance or noise since only one parameter may be identified without error using such a signal. Sinusoidal signals are persistently exciting of second order since their spectra are nonzero at two frequency points. Finally, a random signal can be persistently exciting of any order since its spectrum is nonzero over a frequency interval. It is also noteworthy that a signal periodic with period n can at most be persistently exciting of order n.

Violation of the persistent excitation condition does not mean that obtaining estimates for parameters is impossible. It implies, however, that parameters do not converge to true values no matter how many data points are taken.

7.2.2.4 Frequency-Domain Bias Distribution Under PEM

The discussion of parameter convergence is based on the assumption that there exists a "true" parameter vector. Even when the parameters converge to their "best" values, it is still possible for the model to show significant bias from the true plant model if the model structure used for identification is not rich enough. For example, an FIR model with too few coefficients may differ from the true system significantly even with the best choice of impulse response coefficients. Understanding how the choice of input signal affects distribution of model bias in the frequency domain is important, especially for developing a model for closed-loop control purposes, since accuracy of fit in certain frequency regions (*e.g.*, cross-over frequency region) can be more important than others.

In the prediction error method, parameters are fitted based on the criterion

$$\min_{\theta} \frac{1}{N} \sum_{k=1}^{N} \hat{e}_{pred}^2(k,\theta) \tag{7.42}$$

where $\hat{e}_{pred}(k,\theta) = H^{-1}(q,\theta)\{\hat{y}(k) - G(q,\theta)u(k)\}$. Suppose the true system is represented by

$$\hat{y}(k) = G_o(q)u(k) + H_o(q)\varepsilon(k) \tag{7.43}$$

Then,

$$\hat{e}_{pred}(k,\theta) = \frac{G_o(q) - G(q,\theta)}{H(q,\theta)}u(k) + \frac{H_o(q)}{H(q,\theta)}\varepsilon(k) \tag{7.44}$$

By Parseval's theorem,

$$\begin{aligned} & \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \hat{e}_{pred}^2(k,\theta) && (7.45)\\ = & \int_{-\pi}^{\pi} \Phi_{\hat{e}}(\omega) d\omega && (7.46)\\ = & \int_{-\pi}^{\pi} \left(\left| G_o(e^{j\omega}) - G(e^{j\omega},\theta) \right|^2 \frac{\Phi_u(\omega)}{\left| H(e^{j\omega},\theta) \right|^2} + \frac{\left| H_o(e^{j\omega}) \right|^2}{\left| H(e^{j\omega},\theta) \right|^2} \Phi_{\varepsilon}(\omega) \right) d\omega \end{aligned}$$

where $\Phi_{\hat{e}}(\omega)$ is the spectrum of $\hat{e}_{pred}(k)$.

Note that, in the case that the disturbance model does not contain any unknown parameter,

$$\begin{aligned} & \lim_{N\to\infty} \tfrac{1}{N} \textstyle\sum_{k=1}^{N} \hat{e}_{pred}^2(k,\theta) \\ = & \int_{-\pi}^{\pi} \left(\left| G_o(e^{j\omega}) - G(e^{j\omega},\theta) \right|^2 \frac{\Phi_u(\omega)}{\left| H(e^{j\omega}) \right|^2} + \frac{\left| H_o(e^{j\omega}) \right|^2}{\left| H(e^{j\omega}) \right|^2} \Phi_{\varepsilon}(\omega) \right) d\omega \end{aligned} \tag{7.47}$$

Since the last term of the integrand is unaffected by the choice of θ, we may conclude that PEM selects θ such that the L_2-norm of the error $G_o(q) - G(q,\theta)$ weighted by the filtered input spectrum $\Phi_{u_f}(\omega)$ (where $u_f(k) = H^{-1}(q)u(k)$) is minimized. An implication is that, in order to obtain a good frequency response estimate at a certain frequency region, the filtered input u_f must be designed so that its power is concentrated in

the region. If we want good frequency estimates throughout the entire frequency range, an input signal with a flat spectrum (*e.g.*, a sequence of independent, zero mean random variables) is the best choice.

Frequency domain bias distribtuion can be made more flexible by minimizing the filtered prediction error $\hat{e}_{fpred}$ ($\triangleq L(q)e_{pred}$). In this case,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \hat{e}^2_{fpred}(k,\theta) \tag{7.48}$$

$$= \int_{-\pi}^{\pi} \left(\left| G_o(e^{j\omega}) - G(e^{j\omega},\theta) \right|^2 \frac{\Phi_u(\omega)}{\left|L(e^{j\omega})\right|^2 \left|H(e^{j\omega})\right|^2} + \frac{\left|H_o(e^{j\omega})\right|^2}{\left|L(e^{j\omega})\right|^2 \left|H(e^{j\omega})\right|^2} \Phi_\varepsilon(\omega) \right) d\omega$$

Hence, by prefiltering the data before the parameter estimation, one can affect the bias distribution. This gives an added flexibility when the input spectrum cannot be adjusted freely.

Finally, we have based our argument on the case where the disturbance model does not contain any parameter. When the disturbance model contains some of the parameters, the noise spectrum $|H_o(e^{j\omega})|^2$ does affect the bias distribution. However, the qualitative effects of the input spectrum and prefiltering remain the same.

7.2.3 PARAMETER ESTIMATION VIA STATISTICAL METHODS

In formulating the prediction error minimization, we did not require an exact statistical description of the *underlying plant*. Prediction error minimization is a logical criterion for parametric identification regardless of the true nature of the underlying plant (*i.e.*, even if the assumed model structure does not match the real plant exactly). In stochastic identification, a specific stochastic model is assumed for the underlying plant and plant parameters are estimated in an optimal fashion based on some well-defined