3 Adaptive IIR Filters

Some motivation from system identification theory

- A general prediction model for the system identification (direct filtering) problem is discussed. A characterization is also introduced, mainly that related to a possible optimal solution, i.e., *Maximum Likelihood* estimate.
- A brief discussion of system identification methods as possible candidates to be used in adaptive IIR filters is presented.
- An also a brief review of the more elaborate optimal problem: inverse casual IIR Wiener filtering is included.

- In a different way of what happen with FIR filters, **causal stable Wiener IIR filters to solve the problems of inverse or direct filtering have not a closed form**.
- In particular, inverse Wiener causal IIR filters are more restricted than direct causal IIR filters.
- That is the main reason that in general justify that almost all adaptive IIR filters known are mainly addressed to solve the direct causal filtering problem and not the inverse causal IIR problem.

Characteristics

- FIR Adaptive Filters
  - Stability guaranteed.
  - Unimodality of the performance criterion (MSE).
  - Easy stability conditions for the updating algorithm.
  - High complexity for modeling real world systems.
- IIR Adaptive Filters
  - Low complexity to model real world systems.
  - Multimodality of the performance criterion (MSOE).
  - The IIR adaptive filter requires stability check: exponential stability is not trivial.
  - Stability of the updating algorithm must be guaranteed.

Aspects to study (and improve!)

- Modeling capability.
- Computational complexity.
- Convergence speed.

3.1 System identification and adaptive IIR filtering

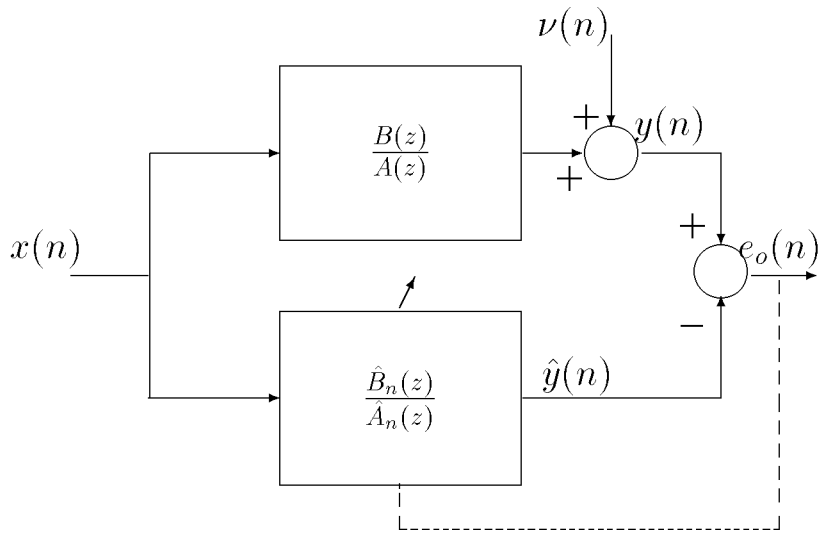

Figure 23: System identification configuration

- In system identification asymptotic properties for the estimators are assumed using a stationary environment and a decreasing convergence factor.
- The models of system identification include noise parameters.
- Lines followed:
  1. Stochastic approximation.
  2. Pseudolinear regression or Model reference methods (Pseudolinear regression, PLR).
  3. Modification of off-line methods (Recursive prediction error RPE, Recursive least squared RLS).

Outline:

- Setup of a general model set that is not function of data: prediction error definition.
- Relationship between Maximum Likelihood (ideal properties pursued) and prediction error optimization.
- Conditions of *Identifiability* of the general model set and the input.

3.1.1 Some simple model descriptions

In system identification the *linear differences equation* (ARMA model) is often used

$$\begin{aligned} y(n) - a_1 y(n-1) - ... - a_N y(n-N) \; &= \\ b_0 x(n) + ... + b_N x(n-N) + \nu(n) & \end{aligned}$$

or

$$A(q)y(n) = B(q)u(n) + \nu(n)$$

then

$$y(n) = \boldsymbol{\theta}^T \phi(n) + \nu(n) = \hat{y}(n/\theta) + \nu(n)$$

where: $\boldsymbol{\theta} = [a_1 \; ... \; a_N \; b_0 \; ... \; b_N]^T$ and $\hat{y}(n/\theta)$ is a one-step prediction of $y(n)$.

If the noise can be modeled, it is possible the utilization of a more complete *ARMAX model*,

$$\begin{aligned} \nu(n) &= C(q)r(n) \\ A(q)y(n) &= B(q)u(n) + C(q)r(n) \\ \boldsymbol{\theta} &= [a_1 \; ... \; a_N \; b_0 \; ... \; b_N \; c_0 \; ... c_M]^T \end{aligned}$$

In general, the more complete model described by an *state variable description*,

$$\begin{aligned} \boldsymbol{x}(n+1) &= \boldsymbol{A}\,\boldsymbol{x}(n) + \boldsymbol{b}\,u(n) + \nu_1(n) \\ y(n) &= \boldsymbol{c}\,\boldsymbol{x}(n) + du(n) + \nu_2(n) \end{aligned}$$

is given by the *Kalman filter*.

3.2 A general prediction model set

- Let $\mathcal{S}$, a *structure of models*,

$$\mathcal{S} : y(n) = G_o(q)x(n) + H_o(q)r(n)$$
$$G_o(q) = \sum_{k=1}^{\infty} g_k q^{-k} \qquad H_o(q) = 1 + \sum_{k=1}^{\infty} h_k q^{-k}$$

- The *model set*,

$$y(n) \;=\; G(q,\theta)x(n) + H(q,\theta)r(n)$$

with $f_r(x,\theta)$ the probability density function (pdf) of $r(n)$ (white noise), and $\boldsymbol{\theta} \in \mathcal{D}_{\mathcal{M}}$ of dimension $2N$.

- Usually the dynamics of the transfer function and the perturbation have different orders.
- For system identification (and adaptive filters) it is important that $G(q,\theta)$ independent of $H(q,\theta)$, i.e.,

$$\boldsymbol{\theta} = [\boldsymbol{\rho} \;\; \boldsymbol{\eta}]^T \qquad \mathcal{D}_{\mathcal{M}} = \mathcal{D}_\rho \times \mathcal{D}_\eta \qquad \boldsymbol{\rho} \in \mathcal{D}_\rho, \;\; \boldsymbol{\eta} \in \mathcal{D}_\eta$$

then

$$\begin{aligned} G(q,\theta) &= G(q,\rho) \\ H(q,\theta) &= H(q,\eta) \end{aligned}$$

3.2.1 Model set construction: prediction

To design the **noise prediction model** some assumptions are necessary,

- Noise model invertibility (condition on the power spectral density of the noise). If $\nu(n) = H(q)r(n)$, then $r(n) = \tilde{H}(q)\nu(n)$ with $\tilde{H}(q) = \frac{1}{H(q)}$
- One-step predictor of $\nu(n)$,

$$\begin{aligned} \nu(n) &= \sum_{k=0}^{\infty} h_k r(n-k) = r(n) + \sum_{k=1}^{\infty} h_k r(n-k) \\ &= r(n) + m(n-1) \end{aligned}$$

  Assuming $\nu(k)$ known for $k \leq n-1$ this is leads to the prediction of $\nu(n)$.

- With $r(n)$ white noise, $P(x \leq r(n) \leq x + \Delta x) \cong f_r(x)\Delta x$. And using the pdf of $\nu(n)$

$$\begin{aligned} f_\nu(x)\Delta x &= P(x \leq \nu(n) \leq x + \Delta x / \nu_{-\infty}^{n-1}) \\ &= P(x \leq r(n) + m(n-1) \leq x + \Delta x) \\ &= P(x - m(n-1) \leq r(n) \leq x + \Delta x - m(n-1)) \\ &= f_r(x - m(n-1))\Delta x \end{aligned}$$

- In general we adopt the (conditional) mean of $f_\nu(x)$, i.e.,

$$\hat{\nu}(n/n-1) = m(n-1) = \sum_{k=1}^{\infty} r(n-k)$$

  note that this estimate minimizes

$$\min_{\tilde{\nu}} E\{(\nu(n) - \tilde{\nu}(n))^2\} = \hat{\nu}(n/n-1)$$

- Finally

$$\hat{\nu}(n/n-1) = [H(q) - 1]\, r(n) = \left[1 - \frac{1}{H(q)}\right] \nu(n)$$

The **output prediction** design follows this steps

- Objective: prediction of $y(n) = G(q)x(n) + \nu(n)$, based on

$$\nu(k) = y(k) - G(q)x(k) \quad k \leq n-1$$

- Then

$$\begin{aligned}
\hat{y}(n/n-1) &= G(q)x(n) + \hat{\nu}(n/n-1) \\
&= G(q)x(n) + \left[1 - \frac{1}{H(q)}\right]\nu(n) \\
&= G(q)x(n) + \left[1 - \frac{1}{H(q)}\right][y(n) - G(q)x(n)] \\
&= \frac{1}{H(q)}G(q)x(n) + \left[1 - \frac{1}{H(q)}\right]y(n)
\end{aligned}$$

or

$$\hat{y}(n/\theta) = \frac{1}{H(q,\theta)}G(q,\theta)x(n) + \left[1 - \frac{1}{H(q,\theta)}\right]y(n)$$

that not depends on $f_r(x,\theta)$.

3.2.2 Model representation: transfer functions

- Associating suitable transfer functions to the model set,

$$A(q)y(n) = \frac{B(q)}{F(q)}x(n) + \frac{C(q)}{D(q)}r(n)$$

- Then, the related predictor is

$$\hat{y}(n/n-1) = \frac{D(q)B(q)}{C(q)F(q)}x(n) + \left[1 - \frac{D(q)A(q)}{C(q)}\right]y(n)$$

$B(q)$	$\rightarrow$	FIR
$A(q), B(q)$	$\rightarrow$	ARX
$A(q), B(q), C(q)$	$\rightarrow$	ARMAX
$A(q), C(q)$	$\rightarrow$	ARMA
$A(q), B(q), D(q)$	$\rightarrow$	ARARX
$A(q), B(q), C(q), D(q)$	$\rightarrow$	ARARMAX
$B(q), F(q)$	$\rightarrow$	output error

Some special cases

- *Equation error model*

$$G(q) = \frac{B(q)}{A(q)} \qquad H(q) = \frac{1}{A(q)}$$

- *Pseudolinear regression model*

$$G(q) = \frac{B(q)}{C(q)} \qquad H(q) = \frac{1 - A(q)}{C(q)}$$

- *Output error model*

$$G(q) = \frac{B(q)}{F(q)} \qquad H(q) = 1$$

3.2.3 Maximum Likelihood and prediction error

- Maximum Likelihood concept: Given a function that depends on the data and the unknown parameters (the Likelihood function), the estimate is associated to the parameter that maximizes this function.

- For independent observations
  - Likelihood function: conditional pdf of the observations (product of partial conditional pdf, known a priori).

- For sequential observations
  - Computation of conditional pdf at time $n+1$ based on observation up to time n. Then, similar to a prediction problem, i.e., the Likelihood function is obtained as a product of conditional pdf in the prediction errors.

- We saw that for prediction error methods: the prediction model and the criteria to be minimized are independent of the statistics of the signal involved. The same result is expected for this method.

- **The Maximum Likelihood method**: Let y a random variable with pdf $p(y, \theta)$, θ unknown. To estimate θ from observation y, choose θ that maximizes $L(\theta, y) = p(y, \theta)$

- Then

  - For independent observations

    $L(\theta, y(1)...y(N)) = p(y(1),\theta)p(y(2),\theta)...p(y(N),\theta)$. For the Gaussian case:

$$p(y,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-m)^2}{2\sigma^2}}$$

    where m and σ depends on θ. Then

$$L(\theta, y(1)...y(N)) = \left(\sigma\sqrt{2\pi}\right)^{-N} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[y(i)-m]^2}$$

    or

$$-\log L = \frac{1}{\sigma^2}\sum_{i=1}^{N} e(i)^2 + N\log\sigma + \frac{N}{2}\log 2\pi$$

    where $e(i) = y(i) - m$. Since maximization of L is equivalent to the minimization of $-\log L$,

    * If σ is (a constant) known, then is equivalent to minimization of $V = \frac{1}{2}\sum_{i=1}^{N} e(i)^2$.

    * If σ is unknown: first minimization of V. Second: estimate $\hat{\sigma}^2 = \frac{2}{N}\min V$.

– For sequential observations

* Let $\{y(i)\}$ a sequence of observations and $\boldsymbol{y}_N$ a vector of the last N.

* Then $L(\theta, \boldsymbol{y}_N) = p(\boldsymbol{y}_N/\theta)$.

* But $p(\boldsymbol{y}_N/\theta) = p(y(N)/\boldsymbol{y}_{N-1}/\theta)p(\boldsymbol{y}_{N-1}/\theta)$.

* Then, in general $L(\theta, \boldsymbol{y}_N) = p(y(N)/\boldsymbol{y}_{N-1})...p(y(2)/y(1))p(y(1))$.

* If $p(y(n)/\boldsymbol{y}_{n-1}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\hat{y})^2}{2\sigma^2}}$, the conditional mean is

$$\hat{y}(\theta, \boldsymbol{y}_{n-1}) \;=\; E\{y(n)/\boldsymbol{y}_{n-1}\}$$

and the conditional covariance

$$\sigma(\theta, \boldsymbol{y}_{n-1}) \;=\; cov\{y(n)/\boldsymbol{y}_{n-1}\}$$

* Finally

$$-\log L(\theta; \boldsymbol{y}_N) \;=\; \frac{1}{2}\sum_{i=1}^{N}\left[(e(i)/\sigma(i))^2 + \log\sigma(i)\right] + \frac{N}{2}\log 2\pi$$

where $e(i) = \begin{cases} y(1) - m & if\ i = 1 \\ y(i) - \hat{y}(i) & i = 2, ...N \end{cases}$.

* If σ is known, this is equivalent to minimization of $V = \frac{1}{2}\sum_{i=1}^{N} e(i)^2$. If $\sigma(i)$ are unknown similar procedure to the case of independent observations.

- Then: interpretation of $\hat{y}(n)$ as the **prediction** of $y(n)$ based on data up to time $n-1$.
- Then, **Prediction error principle**: postulate a model to determine $\hat{y}(n)$ as a function of $y(n)$ and θ. Then adjust θ in such a way that the criteria is minimum.
- **Some properties** of Maximum Likelihood method:
  - For independent observations:
    1. Consistent.
    2. Asymptotically Normal.
    3. Asymptotically efficient.
  - For sequential observations
    1. Characteristics dependent of the known dynamic of the model.
    2. Consistent and efficient for the general ARMAX model.
- Based on the prediction error minimization as general criteria

$$V_n(\theta) \;=\; \frac{1}{N}\sum_{n=1}^{N}\left(y(n)-\hat{y}(n/n-1)\right)^2$$

the obtained estimate $\hat{\rho}$ is independent of η and $H(q,\eta)$, in such form that asymptotically $G(q,\hat{\rho}) \to G_o(q)$.

3.2.4 Model Identifiability: sufficient and insufficient order

- **Definition**: A model structure $\mathcal{M}$ is globally identifiable in θ^* if $\mathcal{M}(\theta) = \mathcal{M}(\theta^*)$, $\theta \in \mathcal{D}_{\mathcal{M}}$, implies that $\theta = \theta^*$ for almost all $\theta^* \in \mathcal{D}_{\mathcal{M}}$.

- In particular, if this condition is satisfied for all $\theta^* \in \mathcal{D}_{\mathcal{M}}$, then $\mathcal{M}$ is globally identifiable in strict from.

- For the proposed general structure of models, $\mathcal{M}$ is globally identifiable in θ^*, if and only if

$$G(z,\theta) = G(z,\theta^*) \quad and \quad H(z,\theta) = H(z,\theta^*)$$

  for almost all z, this implies that $\theta = \theta^*$.

- **Theorem**: For the proposed structure of models, the global condition of Identifiability is guaranteed in θ^* if and only if

  1. No common factors exist in $A(q)$, $B(q)$ and $C(q)$.
  2. No common factors exist in $B(q)$, and $F(q)$.
  3. No common factors exist in $C(q)$, and $D(q)$.
  4. If $n_a \geq 1$, no common factors exist in $F(q)$, and $D(q)$.

For the particular case of the output error model

- Notation: $B(z) = z^{n_b}\overline{B}(z)$, $\overline{B}(z) = \overline{b}_1 z^{n_b-1} + ... + \overline{b}_{n_b}$, $F(z) = z^{n_f}\overline{F}(z)$, $\overline{F}(z) = \overline{f}_1 z^{n_f-1} + ... + \overline{f}_{n_f}$, then the analysis is independent if $B(z)$ or $F(z)$ have zero coefficients b_{n_b} or f_{n_f}.
- Then

$$\begin{aligned} G(z,\theta^*) &= \frac{B^*(z)}{F^*(z)} \\ &= z^{n_f-n_b}\frac{z^{n_b}B^*(z)}{z^{n_f}F^*(z)} \\ &= z^{n_f-n_b}\frac{\overline{B}^*(z)}{\overline{F}^*(z)} \end{aligned}$$

- the following condition must be satisfied

$$\begin{aligned} G(z,\theta^*) &= G(z,\theta) = \frac{B(z)}{F(z)} \\ &= z^{n_f-n_b}\frac{\overline{B}(z)}{\overline{F}(z)} \end{aligned}$$

  or, as consequence

$$\overline{F}(z)\overline{B}^*(z) - \overline{F}^*(z)\overline{B}(z) = 0 \tag{15}$$

- $\overline{F}^*(z)$ is a n_f-order polynomial, then has n_f zeros α_i, such as

$$\overline{F}^*(\alpha_i) = 0$$

  for $i = 1, ..., n_f$.
- If $\overline{F}^*(z)$ and $\overline{B}^*(z)$ are coprime, i.e., $\overline{B}^*(\alpha_i) \neq 0$ for $i = 1, ..., n_f$, then to satisfy (15), the following condition must be verified

$$\overline{F}(\alpha_i) = 0$$

  for $i = 1, ..., n_f$. But this implies that $\overline{B}(z) = \overline{B}^*(z)$, such as finally $\theta = \theta^*$.

- If $\overline{F}^*(z)$ and $\overline{B}^*(z)$ have a common factor such as

$$\begin{aligned} \overline{F}^*(z) &= \gamma(z)\overline{F}_1^*(z) \\ \overline{B}^*(z) &= \gamma(z)\overline{B}_1^*(z) \end{aligned}$$

this leads to

$$\begin{aligned} \overline{F}(z) &= \beta(z)\overline{F}_1^*(z) \\ \overline{B}(z) &= \beta(z)\overline{B}_1^*(z) \end{aligned}$$

for $\gamma(z)$ and $\beta(z)$ arbitrary in (15). Then the output error prediction model is not globally identifiable in θ^*.

- Definition: A structure $\mathcal{M}_1$ is in $\mathcal{M}_2$, $\mathcal{M}_1 \subset \mathcal{M}_2$ if $\mathcal{D}_{\mathcal{M}_1} \subset \mathcal{D}_{\mathcal{M}_2}$, and the mapping is obtained constraining $\mathcal{M}_2$ in order that $\theta \in \mathcal{D}_{\mathcal{M}_1}$.

  Example: $\mathcal{M}_2$: model set of order m, $\mathcal{M}_1$ model set of order n, such as $m < n$, then $\mathcal{M}_1 \subset \mathcal{M}_2$.

- With this definition, the ideal model is in structure $\mathcal{M}$, if $\mathcal{S} \subset \mathcal{M}$, i.e., $n_a \geq n_a^o$, $n_b \geq n_b^o$, etc.

- The conditions of global Identifiability of the previous theorem can be rewritten as

  Theorem: Let $\mathcal{S}$ with n_a^o, n_b^o, etc., polynomial orders (with all common factors cancelled). Then $\mathcal{S} \subset \mathcal{M}$ and θ is globally identifiable, if and only if

  1. $\min(n_a - n_a^o, n_b - n_b^o, n_c - n_c^o) = 0$.
  2. $\min(n_b - n_b^o, n_f - n_f^o) = 0$.
  3. $\min(n_c - n_a^c, n_d - n_d^o) = 0$.
  4. If $n_a > 0$, $\min(n_f - n_f^o, n_d - n_d^o) = 0$.

3.2.5 Model selection: persistent excitation

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ two structures in $\mathcal{M}$. A condition on the input data is that it must allow the discrimination between $\mathcal{M}_1$ and $\mathcal{M}_2$.

Let $\theta_1 \in \mathcal{M}_1$ and $\theta_2 \in \mathcal{M}_2$, with $e_i(n, \theta_i)$ the associated prediction error, and $G_i(q) = G(q, \theta_i)$.

Define

$$\begin{aligned} \Delta G(q) &= G_2(q) - G_1(q) \\ \Delta H(q) &= H_2(q) - H_1(q) \end{aligned}$$

If $\theta_o \in \mathcal{S}$ such as $y(n) = G_o(q)x(n) + H_o(q)e_o(n)$, then

$$\begin{aligned} e_2(n) &= \frac{1}{H_2(q)}\left(G_o(q)x(n) + H_o(q)e_o(n)\right) - \frac{1}{H_2(q)}G_2(q)x(n) \\ &= \frac{1}{H_2(q)}\left[(G_o(q) - G_2(q))x(n) + H_o(q)e_o(n)\right] \end{aligned}$$

Then

$$\begin{aligned} \Delta e(n) &= e_1(n) - e_2(n) \\ &= \frac{\Delta H}{H_1 H_2}y(n) - \left(\frac{H_2G_1 - H_1G_2}{H_1H_2}\right)x(n) \\ &= \frac{\Delta H}{H_1 H_2}(H_2e_2(n) + G_2x(n)) - \left(\frac{H_2G_1 - H_1G_2}{H_1H_2}\right)x(n) \\ &= \frac{\Delta H}{H_1}e_2(n) + \left(\frac{\Delta HG_2 - H_2G_1 + H_1G_2}{H_1H_2}\right)x(n) \\ &= \frac{\Delta H}{H_1 H_2}[(G_o - G_2)x(n) + H_oe_o(n)] + \left(\frac{\Delta HG_2 - H_2G_1 + H_1G_2}{H_1H_2}\right)x(n) \\ &= \frac{1}{H_1}\left[\left\{\Delta G + \frac{\Delta H}{H_2}(G_o - G_2)\right\}x(n) + \left\{\frac{\Delta H}{H_2}H_o\right\}e_o(n)\right] \end{aligned}$$

Let $E\{v^2(n)\} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{V}(e^{jw})dw$ the *power spectrum* of $v(n)$, then

$$\mathcal{Y}(e^{jw}) = |G(e^{jw})|^2\mathcal{X}(e^{jw}) + \lambda|H(e^{jw})|^2$$

where $\lambda = E\{e^2(n)\}$.

In our problem

$$E\{\Delta e^2(n)\} = \frac{1}{2\pi}\int_{-pi}^{\pi} \frac{1}{|H_1|^2}\left[\left|\Delta G + \frac{\Delta H}{H_2}(G_o - G_2)\right|^2 \mathcal{X}(e^{jw}) + |\Delta H|^2 \left|\frac{H_o}{H_2}\right|^2 \lambda_o\right] dw$$

where $\lambda_o = E\{e_o^2(n)\}$.

Because $|H_o(e^{jw})|^2 > 0$ (noise model invertible), we consider the case where $E\{\Delta e^2(n)\} = 0$, but $\Delta G(e^{jw})$ and $\Delta H(e^{jw})$ are not simultaneously zero.

If $\Delta H(e^{jw}) = 0$, then $|\Delta G(e^{jw})|^2 \mathcal{X}(e^{jw}) = 0$. Thus $\Delta G(e^{jw}) = 0$, and as consequence the models are identical (the data allows model discrimination).

Definition: $x(n)$ with power spectral density $\mathcal{X}(e^{jw})$ is **persistent of order** r if for all filter $M_r(q) = m_1 q^{-1} + ... + m_r q^{-r}$,

$$|M(e^{jw})|^2 \mathcal{X}(e^{jw}) = 0 \quad implies \quad M_r(e^{jw}) = 0$$

Because $M(z^{-1})M(z)$ could have $r-1$ zeros in $|z| = 1$, then $x(n)$ is persistent of order r if $\mathcal{X}(e^{jw}) \neq 0$ at least in r points in $-\pi \leq \pi$.

In general $x(n)$ is persistent if $\mathcal{X}(e^{jw}) > 0$.

For the output error model

$$G(z,\theta) = \frac{B(z,\theta)}{F(z,\theta)} \qquad \Delta G(z,\theta) = \frac{B_1(z,\theta)F_2(z,\theta) - B_2(z,\theta)F_1(z,\theta)}{F_1(z,\theta)F_2(z,\theta)}$$

then

$$|B_1F_2 - B_2F_1|^2 \mathcal{X}(e^{jw}) \;=\; 0$$

Since $\Delta G(z)$ has order n_b+n_f-1, then if $x(n)$ is persistent of order n_b+n_f-1, $\Delta G(e^{jw}) = 0$ and the input data is sufficient to allow the discrimination between the two models.

If numerator and denominator are of order r,

- $x(n)$ is persistent of order $2r+1$.
- $\mathcal{X}(e^{jw}) \neq 0$ in $2r+1$ points.
- $x(n)$ can be formed by $r+1$ sinusoids, $x(n) = \sum_{k=1}^{r+1} u_k \cos(w_k n)$, with $w_k \neq w_j$, $k \neq j$, $w_k \neq 0$.

3.3 A first look to recursive methods

3.3.1 Stochastic approximation

Based on the simple model

$$y(n) \;=\; \boldsymbol{\varphi}^T(n)\boldsymbol{\theta} + \nu(n)$$

the objective is to minimize

$$V(\boldsymbol{\theta}) \;=\; \frac{1}{2}E\left\{\left(y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}\right)^2\right\}$$

where $\boldsymbol{\varphi}(n) = [y(n-1)...y(n-n_a)x(n-1)...x(n-n_b)]^T$, $\boldsymbol{\theta} = [a_1...a_{n_a}b_1...b_{n_b}]^T$. Then

$$\left[-\frac{\partial V(\boldsymbol{\theta})}{\partial \theta}\right] \;=\; E\left\{\boldsymbol{\varphi}(n)\left(y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}\right)\right\} = 0$$

(note that if $E\{f(n)\}$ is replaced by $\frac{1}{N}\sum_1^N f(n)$ the Least Squares estimate is obtained).

We search the solution of the following problem. Given

$$E\{\boldsymbol{Q}(\boldsymbol{y}(n), e(n))\} \;=\; 0 \qquad (16)$$

obtain $\boldsymbol{y}(n)$ in such a way that the measured $\boldsymbol{Q}(\boldsymbol{y}, e)$ satisfy (16).

A recursive form of the solution is the following

$$\boldsymbol{y}(n) \;=\; \boldsymbol{y}(n-1) + \gamma(n)\boldsymbol{Q}(\boldsymbol{y}(n-1), e(n))$$

where $\gamma(n) \to 0$ for $n \to \infty$.

Example: mean value estimator

$$\begin{aligned} E\{e(n) - y\} &= 0 \\ y(n) &= y(n-1) + \gamma(n)(e(n) - y(n-1)), \quad with\ \gamma(n) = 1/n \\ x(n) &= \frac{1}{n}\sum_{k=1}^{n} e(k) \end{aligned}$$

Applied in the linear regression problem

$$\boldsymbol{\theta}(n) = \boldsymbol{\theta}(n-1) + \gamma(n)\boldsymbol{\varphi}(n)\left[y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n-1)\right]$$

some frequently used variants

- If $\gamma(n) = \gamma_o$ constant.
- If $\gamma(n) = \gamma_o/|\boldsymbol{\varphi}(n)|^2$, normalization.
- If $\gamma(n) = 1/\sum_{k=1}^{n} |\boldsymbol{\varphi}(k)|^2$, normalized and decrescent.

3.3.2 Pseudolinear regression or Model reference methods

Using $y(n) = \boldsymbol{\varphi}^T(n)\boldsymbol{\theta} + \nu(n)$ some *not observable* variables are introduced in $\boldsymbol{\varphi}(n)$. This leads to more general models.

Example: LS method extended to ARMAX model

$$A(q)y(n) = B(q)x(n) + C(q)r(n)$$

Define: $\boldsymbol{\varphi}_o = [y(n-1)...y(n-n_a)x(n-1)...x(n-n_b)r(n-1)...r(n-n_r)]^T$, $\boldsymbol{\theta}(n) = [a_1...a_{n_a}\ b_1...b_{n_b}\ c_1...c_{n_c}]^T$, then

$$y(n) = \boldsymbol{\varphi}_o^T\boldsymbol{\theta} + r(n)$$

this equation is linear in the parameters, then a possible method is RLS. But $\boldsymbol{\varphi}_o$ has some not observable variables that it is necessary replace for suitable estimates.

Consider the following equality

$$r(n) = A(q)y(n) - B(q)x(n) - [C(q) - 1]r(n)$$

then

$$\epsilon(n) = \hat{A}(q)y(n) - \hat{B}(q)x(n) - [\hat{C}(q) - 1]\epsilon(n)$$

and $\boldsymbol{\varphi}(n) = [y(n-1)...y(n-n_a)x(n-1)...x(n-n_b)\epsilon(n-1)...\epsilon(n-n_r)]^T$, then

$$\epsilon(n) = y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n)$$

is the *a posteriori prediction error*. Then, a complete algorithm is the following

$$\begin{aligned}
\epsilon(n) &= y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n-1) \\
\boldsymbol{\theta}(n) &= \boldsymbol{\theta}(n-1) + \frac{1}{n}\boldsymbol{R}^{-1}(n)\boldsymbol{\varphi}(n)\epsilon(n) \\
\boldsymbol{R}(n) &= \boldsymbol{R}(n-1) + \frac{1}{n}\left[\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n) - \boldsymbol{R}(n-1)\right]
\end{aligned}$$

3.3.3 Modification of off-line methods

For slow time varying linear systems usually a modification of off-line methods is contemplated by the introduction of a **forgetting profile**.

- **Example 1, Recursive Least Squares**: For the FIR case this was introduced in the previous chapter. With suitable definitions for the regressor and parameter vector, the criteria to be minimized is

$$V_N(\theta) \;=\; \sum_{n=1}^{N} \overline{\beta}(N,n)[y(n) - \boldsymbol{\theta}^T(n)\boldsymbol{\varphi}(n)]^2$$

then the direct solution is

$$\boldsymbol{\theta}(N) \;=\; \left[\sum_{n=1}^{N} \overline{\beta}(N,n)\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)\right]^{-1} \left[\sum_{n=1}^{N} \overline{\beta}(N,n)\boldsymbol{\varphi}(n)y(n)\right]$$

This can be rewritten in a recursive form only if $\overline{\beta}(n,k)$ satisfy some constraints

$$\overline{\beta}(n,k) \;=\; \lambda(n)\overline{\beta}(n-1,k) \quad 1 \le k \le n-1$$

or

$$\overline{\beta}(n,k) \;=\; \left[\prod_{j=k+1}^{n} \lambda(j)\right] \alpha(k)$$

where $\overline{\beta}(k,k) = \alpha(k)$. Since in general $\lambda(k) \le 1$, if $\lambda(k) = \lambda$ a constant, then

$$\overline{\beta}(n,k) \;=\; \lambda^{n-k}\alpha(k)$$

Defining $\overline{\boldsymbol{R}}(n) = \sum_{k=1}^{n} \overline{\beta}(n,k)\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)$, then

$$\begin{aligned}
\boldsymbol{\theta}(n) &= \boldsymbol{\theta}(n-1) + \overline{\boldsymbol{R}}^{-1}(n)\boldsymbol{\varphi}(n)\alpha(n)\left[y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n-1)\right] \\
\overline{\boldsymbol{R}}(n) &= \lambda(n)\overline{\boldsymbol{R}}(n-1) + \alpha(n)\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)
\end{aligned}$$

or using the matrix inversion lemma with $\boldsymbol{P}(n) = \overline{\boldsymbol{R}}^{-1}(n)$,

$$\begin{aligned}
\boldsymbol{\theta}(n) &= \boldsymbol{\theta}(n-1) + \boldsymbol{L}(n)\left[y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n-1)\right] \\
\boldsymbol{L}(n) &= \frac{\boldsymbol{P}(n-1)\boldsymbol{\varphi}(n)}{\frac{\lambda(n)}{\alpha(n)} + \boldsymbol{\varphi}^T(n)\boldsymbol{P}(n-1)\boldsymbol{\varphi}(n)} = \alpha(n)\boldsymbol{P}(n)\boldsymbol{\varphi}(n) \\
\boldsymbol{P}(n) &= \frac{1}{\lambda(n)}\left[\boldsymbol{P}(n-1) + \frac{\boldsymbol{P}(n-1)\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)\boldsymbol{P}(n-1)}{\frac{\lambda(n)}{\alpha(n)} + \boldsymbol{\varphi}^T(n)\boldsymbol{P}(n-1)\boldsymbol{\varphi}(n)}\right]
\end{aligned}$$

Note that assuming a linear regression model $y(n) = \boldsymbol{\theta}_o\boldsymbol{\varphi}(n) + \nu(n)$, the Least Squared estimate is

$$\begin{aligned}
\boldsymbol{\theta}(N) &= \left[\sum_{n=1}^{N} \overline{\beta}(N,n)\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)\right]^{-1}\left[\sum_{n=1}^{N} \overline{\beta}(N,n)\{\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)\boldsymbol{\theta}_o + \boldsymbol{\varphi}(n)\nu(n)\}\right] \\
&= \boldsymbol{\theta}_o + \left[\sum_{n=1}^{N} \overline{\beta}(N,n)\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)\right]^{-1}\left[\sum_{n=1}^{N} \overline{\beta}(N,n)\{\boldsymbol{\varphi}(n)\nu(n)\}\right]
\end{aligned}$$

then for $N \to \infty$, $\boldsymbol{\theta}(N) \to \boldsymbol{\theta}_o$ only if

- $\nu(n)$ is not correlated with the regressor.
- $\nu(n)$ is not correlated with the input.

then in general, LS or RLS, are **biased**.

- **Example 2, the stochastic approximation**: There is necessary a choice of $\gamma(n)$ in

$$\begin{aligned}\boldsymbol{\theta}(n) &= \boldsymbol{\theta}(n-1) + \gamma(n)\boldsymbol{R}^{-1}(n)\boldsymbol{\varphi}(n)\left[y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n-1)\right] \\ \boldsymbol{R}(n) &= \boldsymbol{R}(n-1) + \gamma(n)\left[\boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n) - \boldsymbol{R}(n-1)\right]\end{aligned}$$

The usual restrictions

- $\gamma(n) \geq 0$.
- $\sum_1^\infty \gamma(n) = \infty$.
- $\sum_1^\infty \gamma^2(n) < \infty$ or $\gamma(n) \rightarrow \gamma_o$ (small) for $n \rightarrow \infty$.

That is related with forgetting factor. Considering $\overline{\boldsymbol{R}}(n) = \frac{1}{\gamma(n)}\boldsymbol{R}(n)$,

$$\begin{aligned}\boldsymbol{\theta}(n) &= \boldsymbol{\theta}(n-1) + \overline{\boldsymbol{R}}^{-1}(n)\boldsymbol{\varphi}(n)\left[y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n-1)\right] \\ \overline{\boldsymbol{R}}(n) &= \frac{\gamma(n-1)}{\gamma(n)}(1-\gamma(n))\overline{\boldsymbol{R}}(n-1) + \boldsymbol{\varphi}(n)\boldsymbol{\varphi}^T(n)\end{aligned}$$

then if $\lambda(n) = \frac{\gamma(n-1)}{\gamma(n)}(1-\gamma(n))$ and $\alpha(n) = 1$ the gain of stochastic approximation and the forgetting factor play the same role. In this case,

$$\overline{\beta}(N,n) = \prod_{k=n+1}^{N} \frac{\gamma(k-1)}{\gamma(k)}(1-\gamma(k)) = \frac{\gamma(n)}{\gamma(N)} \prod_{k=n+1}^{N}(1-\gamma(k))$$

- **Example 3, the prediction error method**: Consider the following ARMAX model

$$y(n) + a\, y(n-1) \;=\; b\, u(n-1) + e(n) + c\, e(n-1) \qquad (17)$$

then $\boldsymbol{\theta} = [a\ b\ c]^T$. Then it is necessary to build the predictor $\hat{y}(n/\boldsymbol{\theta})$ of $y(n)$ based on $y(s), u(s)\ 0 \leq s \leq n-1$. Then, rewriting (17), we obtain

$$\hat{y}(n/\boldsymbol{\theta}) \;=\; -a\, y(n-1) + b\, u(n-1) + c\, \hat{e}(n-1) \qquad (18)$$

where $\hat{e}(n)$ can be obtained recursively from

$$\hat{e}(s) \;=\; y(s) + a\, y(s-1) - b\, u(s-1) - c\, \hat{e}(s-1) \quad s = 1, ..., n \qquad (19)$$

then

$$\hat{y}(n/\boldsymbol{\theta}) + c\, \hat{y}(n-1/\boldsymbol{\theta}) \;=\; (c-a)\, y(n-1) + b\, u(n-1) \qquad (20)$$

The function to minimize is

$$V_n(\boldsymbol{\theta}) \;=\; \frac{1}{2}\sum_{k=1}^{n} \epsilon^2(k, \boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{n} (y(k) - \hat{y}(k/\boldsymbol{\theta}))^2$$

a nonlinear function of the parameter vector $\boldsymbol{\theta}$. Then only an approximated solution can be obtained.

Using a Taylor expansion on $\boldsymbol{\theta}(n-1)$,

$$\begin{aligned} V_n(\boldsymbol{\theta}) &= V_n(\boldsymbol{\theta}(n-1)) + V_n'(\boldsymbol{\theta}(n-1))[\boldsymbol{\theta} - \boldsymbol{\theta}(n-1)] \\ &\quad + \frac{1}{2}[\boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1)]^T V_n''(\boldsymbol{\theta}(n-1))[\boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1)] \\ &\quad + O(\|\boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1)\|^2) \end{aligned}$$

where $O(\|x\|)/\|x\|^2 \rightarrow 0$ for $\|x\| \rightarrow 0$, we obtain

$$\begin{aligned} \boldsymbol{\theta}(n) &= \boldsymbol{\theta}(n-1) - [V_N''(\boldsymbol{\theta}(n-1))]^{-1} [V_N'(\boldsymbol{\theta}(n-1))]^T \\ &\quad + O(\|\boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1)\|^2) \end{aligned}$$

Defining $\psi(n, \boldsymbol{\theta}) = [-\frac{d}{d\boldsymbol{\theta}} \epsilon(n, \boldsymbol{\theta})]$, then

$$\begin{aligned} [V_n'(\boldsymbol{\theta}(n))]^T &= -\sum_{k=1}^{n} \psi(k, \boldsymbol{\theta}) \epsilon(k, \boldsymbol{\theta}) = [V_{n-1}'(\boldsymbol{\theta})]^T - \psi(n, \boldsymbol{\theta}) \epsilon(n, \boldsymbol{\theta}) \\ V_n''(\boldsymbol{\theta}) &= V_{n-1}''(\boldsymbol{\theta}) + \psi(n, \boldsymbol{\theta}) \psi^T(n, \boldsymbol{\theta}) + \epsilon''(n, \boldsymbol{\theta}) \epsilon(n, \boldsymbol{\theta}) \end{aligned}$$

Additional assumptions are necessary at this point,

- $O(\|\boldsymbol{\theta}(n) - \boldsymbol{\theta}(n-1)\|^2) \cong 0$.
- $V_n''(\boldsymbol{\theta}) = V_n''(\boldsymbol{\theta}(n-1))$.
- $\boldsymbol{\theta}(n-1)$ optimum at $n-1$, then $V_{n-1}'(\boldsymbol{\theta}(n-1)) = 0$.
- $\epsilon''(n, \boldsymbol{\theta}(n-1)) \epsilon(n, \boldsymbol{\theta}(n-1)) \cong 0$.

Thus we can write

$$\boldsymbol{R}(n) = \boldsymbol{R}(n-1) + \boldsymbol{\psi}(n, \boldsymbol{\theta}(n-1)) \boldsymbol{\psi}^T(n, \boldsymbol{\theta}(n-1))$$

in such a way that

$$\boldsymbol{\theta}(n) = \boldsymbol{\theta}(n-1) + \boldsymbol{R}^{-1}(n) \boldsymbol{\psi}(n, \boldsymbol{\theta}(n-1)) \epsilon(n, \boldsymbol{\theta}(n-1))$$

To conclude, it remains the computation of $\epsilon(n, \boldsymbol{\theta})$ and $\psi(n, \boldsymbol{\theta})$ from the prediction model (20)

$$\begin{aligned} \frac{d}{da}\hat{y}(n/\theta) + c\,\frac{d}{da}\hat{y}(n-1/\theta) &= -\,y(n-1) \\ \frac{d}{db}\hat{y}(n/\theta) + c\,\frac{d}{db}\hat{y}(n-1/\theta) &= u(n-1) \\ \frac{d}{dc}\hat{y}(n/\theta) + c\,\frac{d}{dc}\hat{y}(n-1/\theta) &= \epsilon(n-1,\theta) \end{aligned}$$

or $\boldsymbol{\psi}(n, \boldsymbol{\theta}) + c\boldsymbol{\psi}(n-1, \boldsymbol{\theta}) = \begin{bmatrix} -y(n-1) \\ u(n-1) \\ \epsilon(n-1,\theta) \end{bmatrix}$.

In order to obtain a recursive gradient computation assume that

- $\epsilon(n, \boldsymbol{\theta}(n-1)) \cong \epsilon(n)$. Then $\epsilon(n) = y(n) - \hat{y}(n)$, with $\hat{y}(n) = -\hat{c}\hat{y}(n-1 + (\hat{c}(n-1) - \hat{a}(n-1))\, y(n-1) + b\, u(n-1)$.
- If $\boldsymbol{\varphi}(n) = \begin{bmatrix} -y(n-1) \\ u(n-1) \\ \epsilon(n-1) \end{bmatrix}$, then $\epsilon(n) = y(n) - \boldsymbol{\varphi}^T(n)\boldsymbol{\theta}(n-1)$.
- If $\boldsymbol{\psi}(n, \boldsymbol{\theta}(n-1)) \cong \boldsymbol{\psi}(n)$, then $\boldsymbol{\psi}(n) = -\hat{c}(n-1)\boldsymbol{\psi}(n-1) + \boldsymbol{\varphi}(n)$

Finally

$$\begin{aligned} \boldsymbol{\theta}(n) &= \boldsymbol{\theta}(n-1) + \frac{1}{n}\boldsymbol{R}^{-1}(n)\boldsymbol{\psi}(n)\epsilon(n) \\ \boldsymbol{R}(n) &= \boldsymbol{R}(n-1) + \frac{1}{n}[\boldsymbol{\psi}(n)\boldsymbol{\psi}^T(n) - \boldsymbol{R}(n-1) \end{aligned}$$

Main characteristics

- Similar to RLS.
- Useful with other models.
- Only local behavior guaranteed.

3.3.4 A first classification: filtered regressor or filtered error algorithms

The main function **Mean square output error** (MSOE), using $\boldsymbol{\theta} = [f_1...f_{n_f}\ b_1...b_{n_b}]^T$, i.e., the output error model and $\boldsymbol{\varphi}(n) = [\hat{y}(n-1)...\hat{y}(n-n_f)x(n-1)...x(n-n_b)]^T$, with

$$\hat{y}(n) = \frac{\hat{B}(q)}{\hat{F}(q)}x(n)$$

this leads to

$$\min_{\theta} V(\theta) = E\{|y(n) - \hat{y}(n)|^2\}$$

that results in general in a non linear problem without closed form solution.

Other useful models to solve the MSOE minimization problem are:

- *Equation error model*

$$G(q) = \frac{B(q)}{A(q)} \qquad H(q) = \frac{1}{A(q)}$$

- *Pseudolinear regression model*

$$G(q) = \frac{B(q)}{C(q)} \qquad H(q) = \frac{1 - A(q)}{C(q)}$$

This leads in general to two different philosophy associated to the MSOE problems solutions

- Filtered regressor - filtered error.
- OE (non linear in the parameters) - EE (linear in the parameters).
- Approximation theory - Stability theory.

3.4 Causal inverse Wiener IIR filtering

- Inverse causal Wiener filters are obtained by minimizing mean square error criterion.
- Realizable inverse IIR Wiener filters, based on IIR-models, are conceptually easy to derive, but explicit solutions for general criteria have been difficult to obtain.
- The *polynomial approach* and the *factorization approach* are used here.
- The causal operation $\{\}_+$, is indirectly evaluated by means of a *Diophantine equation* (polynomial algebraic equation).

Consider the problem model as described by

$$\begin{aligned} y(n) &= G(q)x(n) + H(q)r(n) \\ x(n) &= K(q)u(n) \end{aligned}$$

with $G(q) = \frac{B(q)}{F(q)}$, $H(q) = \frac{C(q)}{D(q)}$ and $\lambda_r = E\{r^2(n)\}$ as usually, but now $K(q) = \frac{E(q)}{A(q)}$ with $\lambda_u = E\{u^2(n)\}$ and $\rho = \lambda_r/\lambda_u$.

The main objective is to estimate $x(n)$ based on data $y(n)$ by minimization of a mean square criteria, i.e.

$$J = E\{(x(n) - \hat{x}(n))^2\} = E\{(x(n) - P(q)y(n))^2\}$$

where $P(q)$ is a stable rational linear operator.

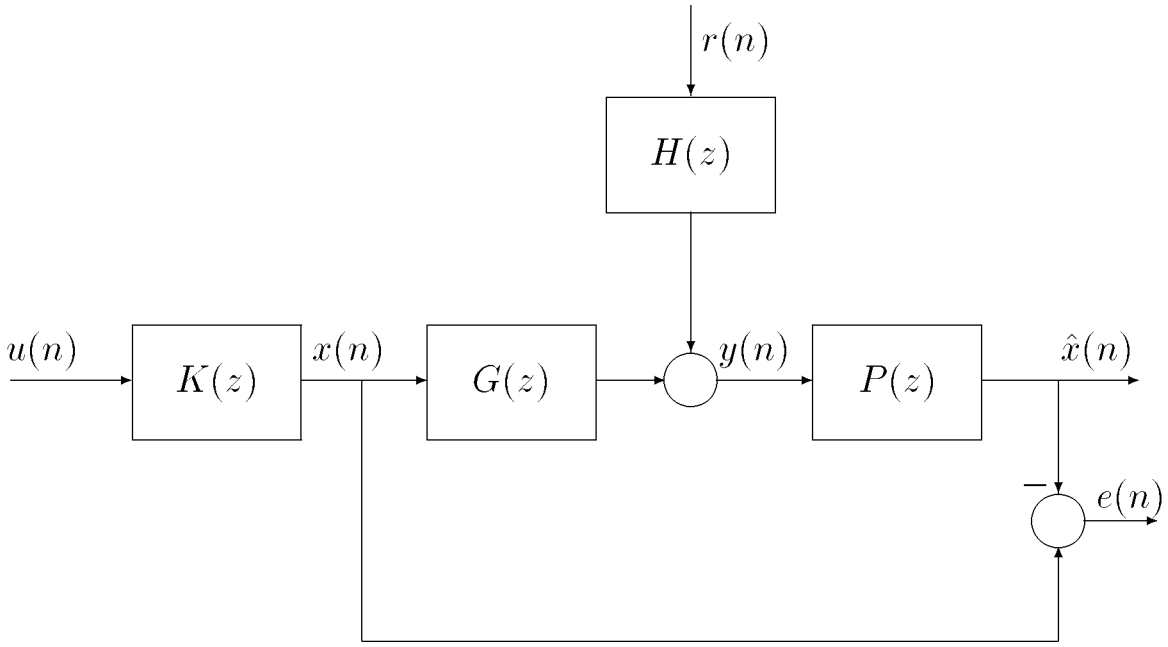

Figure 24: General model for the MSE deconvolution problem

3.4.1 The MMSE deconvolution problem

- Consider the following polynomial spectral factorization

$$\alpha\beta\beta^* = EBD(EBD)^* + \rho CFA(CFA)^*$$

where α is a constant, $\beta(q)$ is the stable and monic spectral factor and $(.)^*$ is the conjugate operator.

- This polynomial also appears as the numerator of the prediction error (innovation) model of $y(n)$, i.e.,

$$y(n) = \frac{\beta}{AFD}\epsilon(n)$$

- A stable $\beta(z)$ exist if and only if $E(z)B(z)$ and $\rho C(z)$ have no common factors with zeros on $|z| = 1$ (system and noise have no common factors).

- The MSE criterion is minimized by the estimator

$$P(z) = \frac{Q_1(z)D(z)F(z)}{\beta(z)}$$

where $Q_1(z)$ together with $L^*(z)$ is the unique solution to the **Diophantine equation** given by

$$z^{-(k+1)}EE^*B^*D^* = Q_1\alpha\beta^* + zAL^*$$

Remarks

- The Diophantine equation can be transformed into an equation in z^{-1} by multiplying both sides by z^{-n_L-1}. If $A(z)$ has zeros in $|z| \leq 1$, this equation will always have a unique solution.

- The optimal filter will have zeros in the pole locations of the measurement noise and of the system, i.e., zeros of $F(z)$ and $D(z)$.

3.4.2 The inner-outer factorization solution

- A stable rational matrix $\boldsymbol{G}^{n|m}(z)$, with full rank $p = \min(m, n)$ for $|z| = 1$ (no zeros on the unit circle), has an **inner-outer factorization**

$$\boldsymbol{G}^{n|m} = \boldsymbol{G}_i^{n|p}\boldsymbol{G}_o^{p|m}$$

with the outer factor $\boldsymbol{G}_o$ having a stable right inverse. It also has a **co-inner-outer factorization**

$$\boldsymbol{G}^{n|m} = \boldsymbol{G}_{co}^{n|p}\boldsymbol{G}_{ci}^{p|m}$$

with the co-outer factor $\boldsymbol{G}_{co}$ having a stable left inverse. If $n \leq m$, the co-outer matrix is square, and its inverse is unique.

Remarks:

- Inner and co-inner matrices are generalizations of scalar all-pass. Multiplication by a (co)inner matrix does not affect the spectral density or power of a signal vector (note that $\boldsymbol{G}$ is co-inner (co-outer), if $\boldsymbol{G}^T$ is inner (outer).
- The important property of outer and co-outer matrices is that they are **stably invertible** (also, the inverses are **causal** if $\boldsymbol{G}_o(0)$ and $\boldsymbol{G}_{co}(0)$ have full rank p).

- Consider now the error $e(n)$ of the figure, that can be written as

$$e(n) \;=\; x(n) - \hat{x}(n) = (1 - P(q)G(q))\,K(q)u(n) - P(q)H(q)r(n)$$

- This can be interpreted in a two-input one-output system as (using the Parseval theorem to work in the frequency domain)

$$\begin{aligned} E\{e(n)e^T(n)\} &= \frac{1}{2\pi j}\,tr\int_{|z|=1} E\{e(z)e^*(z)\}\frac{dz}{z} \\ &= \frac{1}{2\pi j}\,tr\int_{|z|=1}\begin{bmatrix}(1-PG)K\\ PH\end{bmatrix}^T\begin{bmatrix}(1-PG)K\\ PH\end{bmatrix}^*\frac{dz}{z} \end{aligned}$$

- or in terms of a performance index

$$\begin{aligned} J &= \|\,[(1 - P(z)G(z))K(z) \;\; P(z)H(z)]\,\|_2^2 \\ &= \|\,[K(z)\;\;0] \;-\; P(z)[G(z)K(z)\;\;H(z)]\,\|_2^2 \end{aligned}$$

  then, by the factorization of the second term

$$\boldsymbol{U} \;=\; [G(z)K(z)\;\;H(z)] = \boldsymbol{U}_{co}\boldsymbol{U}_{ci} \tag{21}$$

  where $\boldsymbol{U}_{co}$ is co-outer of dimension 1×1 and $\boldsymbol{U}_{ci}$ is co-inner of dimension 1×2.

- The scalar co-outer will have a stable inverse if the left hand side of (21) has full rank 1 $\forall|z| = 1$ (i.e., BE and C have no common factors with zeros on $|z| = 1$).

- The inverse $\boldsymbol{U}_{co}^{-1}(z)$ is causal if and only if $\boldsymbol{U}_{co}(0) \neq 0$.

- Then, premultiplying by $\boldsymbol{U}_{ci}^*$ (norm is preserved) the solution given by

$$P(q) \;=\; \left\{\begin{bmatrix}\dfrac{E(z)}{A(z)} & 0\end{bmatrix}\boldsymbol{U}_{ci}^*\right\}_+ \boldsymbol{U}_{co}^{-1}$$

- The inverse $\boldsymbol{U}_{co}^{-1}$ is guaranteed to be stable (that justify the factorization).

4 Some useful tools

Concepts on Approximation and Stability theory

---

1. Considerations on time variant linear systems.
2. ODE, conditions for the association, Liapunov function. Stationary points: theorems.
3. Approximation concepts: Orthonormal space decomposition of $\mathcal{L}_2$ (interpolation), relationship with Hankel norm.
4. Stability concepts: Stability of a quasi-time-invariant linear system. Stability of a particular non linear system: passivity and hyperstability.