IMPERIAL COLLEGE LONDON

EE3-08

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2012

EEE/ISE PART III/IV: MEng, BEng and ACGI

ADVANCED SIGNAL PROCESSING

Tuesday, 1 May 2:30 pm

Time allowed: 3:00 hours

There are FIVE questions on this paper.

Answer TWO of questions 1, 2, 3 and ONE of questions 4, 5.

All questions carry equal marks

Any special instructions for invigilators and information for candidates are on page 1.

Examiners responsible First Marker(s) : D.P. Mandic, D.P. Mandic
Second Marker(s) : P.T. Stathaki, P.T. Stathaki

1) Consider the first order Markov (that is AR(1)) process described by

$$x[n] = \beta x[n-1] + (1-\beta)w[n]$$

where $w[n] \sim \mathcal{N}(0,1)$, and the coefficient $\beta \in (0,1)$ is chosen so as to provide a convex combination of the past input and the current driving noise sample.

a) Show that the power of this process is given by [4]

$$P = \frac{1-\beta}{1+\beta}$$

and that

$$\beta = \frac{E\{x[n]x[n-1]\}}{E\{x^2[n]\}}$$

(Hint: signal power $P_x = \sigma_x^2 = E\{x^2(n)\}$, noise power $P_w = \sigma_w^2 = E\{w^2(n)\}$)

b) Show that the normalised autocorrelation coefficient [4]

$$\rho_1 = \frac{E\{x[n]x[n-1]\}}{E\{x^2[n]\}} = \frac{r_{xx}(1)}{r_{xx}(0)} = \beta$$

c) Given the first order predictor [6]

$$\hat{x}[n] = h\,x[n-1]$$

where h is the filter coefficient, evaluate the mean square error (MSE) as a function of h, and find the optimal predictor coefficient h_{opt} and the minimal MSE J_{min}. The prediction error is given by

$$e[n] = x[n] - h\,x[n-1]$$

Explain whether the prediction error is orthogonal to $x[n-1]$.

d) Explain how you would find the correct order of an AR(p) process and the concept of partial correlation. Sketch a partial correlation graph for a general AR(1) process. [6]

2) The estimated autocorrelation sequence of a random process $x(n)$ is

$r(0) = 1$, $r(1) = 0.5$, $r(2) = 0.5$, $r(3) = 0.25$, $r(4) = 0.2$, $r(5) = 0.12$, $r(6) = 0.08$.

a) If $x(n)$ is an AR(2) process:

i) Write down the mathematical expression for such a process. What are the requirements for the driving noise of such processes? From the shape of the correlation function, can you tell whether the poles of the AR(2) system are real or complex?

ii) Using the Yule-Walker equations or otherwise find the coefficients a_1 and a_2 of this system. What is the power of the driving noise $w(n)$ for the AR(2) process above? [3]

iii) Assuming that the correlations $r(0), r(1), r(2)$ are accurate and all the others are corrupted by noise, write down a recursive expression for calculating the autocorrelation values for any lag, and calculate the true values for $r(3), r(4), r(5)$. [3]

b) Assume that $x(n)$ is an AR(1) process, and only the correlations $r(0), r(1)$ are not corrupted by noise. Write down the expression for such a process and the true values of correlations $r(2), r(3), r(4), r(5)$. [3]

c) The AR(2) process $x(n)$ from Part a) is observed through a real world sensor so that the measured signal is [3]

$$y(n) = x(n) + q(n), \qquad q \in \mathcal{N}(0, 1)$$

i) Write down the values for the first four autocorrelation values (lags 0 to 3) for the process $y(n)$ if $q(n)$ is uncorrelated with the driving noise $w(n)$ $(q \perp w)$. [2]

ii) Establish the relation between the true AR(2) coefficients of $x(n)$ and those calculated from the noisy process $y(n)$. [3]

iii) Knowing that the noise $q(n)$ is white, how you would calculate the true AR(2) coefficients from the autocorrelation sequence of $y(n)$, without having to incorporate the noise statistics for $q(n)$. [3]

3) Consider the problem of minimum variance unbiased (MVU) estimation.

a) Define the notion of the curvature of the probability density function and explain why it is advantageous to use the *log-likelihood* function instead of the *probability density function.* [4]

b) Consider the Best Linear Unbiased Estimator (BLUE) of a vector parameter.

i) Write down the system model for the BLUE estimation of the unknown vector parameter $\boldsymbol{\Theta} = [\theta_1, \ldots, \theta_p]^T$. State the constraints involved in such estimation. [4]

ii) State the Gauss-Markov theorem. [3]

iii) For the estimation of a constant vector-valued unknown parameter in white Gaussian noise, explain whether you would prefer BLUE, maximum likelihood estimation, or the method of least squares. [3]

c) In a wind farm, the wind speed h is recorded by the control system every 50ms. Over the period of 1s, the measurements $\{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_{20}\}$ are averaged to obtain the estimated wind speed $\hat{h}$. Assume that the expected value $E\{\hat{h}_i\} = ah$ for some constant a, and that due to the stochastic nature of wind, the measurements $\hat{h}_i$ are mutually uncorrelated, that is, $\hat{h}_1 \perp \hat{h}_2 \perp \cdots \perp \hat{h}_{20}$.

If all the measurements have equal variances, $var(\hat{h}_i) = \sigma_h^2$, for $i = 1, \ldots, 20$, determine whether averaging over all the available measurements improves the estimator (in terms of the bias and variance), and sketch the corresponding probability density functions, for the cases [6]

i) $a = 1$ ii) $a = \frac{1}{2}$

4) Consider the class of least squares methods.

a) State the optimisation criterion (cost function) and explain the disadvantages of the method of least squares. How does it deal with non-zero mean signals and nonlinear signal models? [4]

b) Consider the sequential least squares model.

i) Explain the need for the sequential version of the method of least squares and elaborate on its mode of operation. [2]

ii) Derive the sequential least squares solution for the estimation of the DC level in white Gaussian noise. [3]

iii) Derive the expression for the minimum mean squared error J_{min} for the sequential least squares method. [3]

c) The data model is given by (an autoregressive model)

$$x(n) = a_1 x(n-1) + a_2 x(n-2) + \cdots + a_p x(n-p) + w(n) \quad w \sim \mathcal{N}(0, \sigma_w^2)$$

We would like to find the coefficients of this AR(p) process by minimizing a measure of the squared error between the estimated signal

$$\hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k)$$

and the original signal $x(n)$.

i) Using the method of least squares find the expression for the autoregressive coefficients $a_i, i = 1, \ldots, p$. Compare with the Yule-Walker solution. [4]

ii) Explain how an adaptive filter would provide a sequential solution to the problem of finding the AR coefficients. What is the difference in the cost functions of the adaptive filters and the least squares methods? Explain how this difference affects the bias and variance of the AR coefficient estimates. [4]

5) Consider the problem of adaptive filtering of real world signals.

a) Explain in your own words a general principle of the convergence of the unknown parameter vector $\mathbf{w}(n)$ (filter weights) to their optimum values $\mathbf{w}_o$ within the framework of general estimation theory, that is, the meaning of the bias and variance of such an estimator. [4]

b) Derive the optimum weight vector of the Wiener filter in the vector-matrix form. How does that analysis simplify if the filter is operating in the system identification setting and the teaching signal is given by

$$d(n) = \mathbf{w}_o^T \mathbf{x}(n) + q(n) \qquad q \sim \mathcal{N}(0, \sigma_q^2)$$

What is the minimum achievable mean square error (MSE) in this case? [6]

c) The discrete Fourier tranform (DFT) of a sequence $x(n)$ $(n = 0, \ldots, N-1)$ is given by

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\jmath \frac{2\pi}{N} nk} = \sum_{n=0}^{N-1} x(n) W_N^{nk} \qquad \Leftrightarrow \qquad \mathbf{X} = \mathbf{W}\mathbf{x}$$

where $\mathbf{x} = [x(0), \ldots, x(N-1)]^T$, $\mathbf{X}$ is a vector containing the DFT coefficients $\{X(k)\}$ and $\mathbf{W}$ is a matrix of coefficients W_N^{nk}.

i) Write down the expression for the "measurement" matrix $\mathbf{W}$, and explain its role. [4]

ii) Solve this system to obtain the inverse Fourier Transform (iDFT) in the form [3]

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{\jmath \frac{2\pi}{N} kn}$$

(Hint: Due to the orthogonality of complex sinusoidals, $\mathbf{W}^{-1} = \frac{1}{N}\mathbf{W}^H$)

iii) If the data are noisy, explain in your own words how you would perform both Wiener-like and adaptive estimation of the inverse Fourier transform. [3]

Master Copy

EE3-0X

Advanced Signal Processing 2012

1/11

Solutions

1) [**Bookwork and practical application of bookwork**]
a) From the square of the signal

$$x^2[n] = (1-\beta)^2 w^2[n] + 2\beta w[n]x[n-1] + \beta^2 x^2[n-1]$$

and since the driving noise $w[n]$ is independent of $x[n-1]$) which is built up with the past $w[n-j]$), and given that $E\{w^2[n]\} = 1$, we immediately have (since $P = \sigma_x^2 = E\{x^2(n)\} = E\{x^2(n-1)\}$ and $\sigma_w^2 = 1$)

$$\begin{aligned} E\{x^2[n]\} &= (1-\beta)^2 E\{w^2(n)\} + \beta^2 E\{x^2[n-1]\} \\ (1-\beta^2)P &= (1-\beta)^2 \\ \Rightarrow P &= \frac{1-\beta}{1+\beta} \end{aligned}$$

which proves the power relation.
b) To show the relationship with the $r_{xx}(1)$, start from the ACF for lag 1, and calculate

$$\begin{aligned} x[n]x[n-1] &= (1-\beta)w[n]w[n-1] + \beta x^2[n-1] \\ \Rightarrow r_{xx}(1) &= \frac{\beta P}{P} = \beta \end{aligned}$$

c) Since the error $e[n]$ is a function of the unknown parameter h, we can write

$$\begin{aligned} E\{e^2(h)\} &= E\{x^2[n]\} - 2hE\{x[n]x[n-1]\} + h^2 E\{x^2[n-1]\} \\ &= P(1 - 2h\beta + h^2) \\ \Rightarrow h_{opt} &= \beta \end{aligned}$$

We obtain the minimum MSE for h_{opt} and thus

$$J_{min} = J(h_{opt}) = P(1-\beta^2) = (1-\beta)^2$$

For h_{opt} the error is orthogonal to $x[n-1]$ as it would be equal to $w[n]$, which is independent of the data.
d) ACF of $AR(p)$ infinite in duration, **but** can by be described in terms of p nonzero functions ACFs. Denote by a_{kj} the jth coefficient in an autoregressive representation of order k, so that a_{kk} is the last coefficient. Then

$$\rho_j = a_{kj}\rho_{j-1} + \cdots + a_{k(k-1)}\rho_{j-k+1} + a_{kk}\rho_{j-k} \qquad j = 1, 2, \ldots, k$$

leading to the Yule–Walker equation, which can be written as

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_{k1} \\ a_{k2} \\ \vdots \\ a_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix}$$

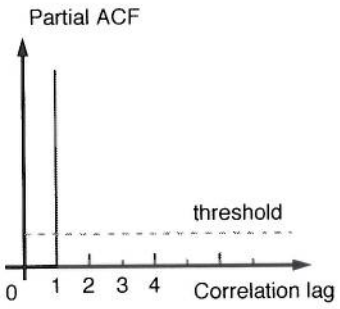

Figure 1: The partial autocorrelation function for an AR(1) process

The partial correlation of an AR(1) process would have only one significant term, at the position 1, like in Figure 1.

2) a) [**bookwork and new examples**]
i) For the AR(2) process we have

$$x(n) = a_1 x(n-1) + a_2 x(n-2) + w(n)$$

where $w(n)$ is the white driving noise (any distribution). The autocorrelation sequence is monotonically decreasing, without a change in sign, so the process is very likely to be low pass with real coefficients a_1 and a_2. This can also be verified from the stability triangle.
ii) The coefficients of an AR(2) model are found by solving the normal equations

$$\begin{bmatrix} r_x(0) & r_x(1) \\ r_x(1) & r_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_x(1) \\ r_x(2) \end{bmatrix}$$

For the given autocorrelation sequence, these become

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

Thus, the coefficients are

$$a_1 = a_2 = \frac{1}{3}$$

Since

$$r(0) = a_1 r(1) + a_2 r(2) + \sigma_w^2 = 1$$

we have

$$\sigma_w^2 = 1 - \frac{1}{3} \times 0.5 - \frac{1}{3} \times 0.5 = \frac{2}{3}$$

iii) The autocorelation sequence for any AR(2) process for lags $k > 0$ is given by

$$r(k) = a_1 r(k-1) + a_2 r(k-2)$$

Thus, for example

$$r(3) = a_1 r(2) + a_2 r(3) \quad r(4) = a_1 r(3) + a_2 r(2) \quad \ldots$$

giving $r(3) = 1/3, r(4) = 0.2778, r(5) = 0.2037$.
b) For an AR(1) process we have

$$x(n) = a_1 x(n-1) + w(n)$$

for which

$$r(k) = a_1 r(k-1) = a_1^k r(0) \qquad k > 0$$

The coefficient a_1 can be found from

$$a_1 = \frac{r(1)}{r(0)} = 0.5$$

The true values for the correlation sequence would hence be

$$r(k) = 0.5^k \quad k > 0 \quad \text{giving} \quad r(2) = 0.25, r(3) = 0.125, r(4) = 0.0625$$

c) If $q \perp w$ we have
i) Since the measurement noise $q[n]$ is white, and $E\{q^2(n)\} = \sigma_q^2 = 1$, we have

$$\begin{aligned} r_{yy}(0) &= E\{y[n]y[n]\} = E\{(x[n] + q[n])(x[n] + q[n])\} = r_{xx}(0) + \sigma_q^2 \\ r_{yy}(1) &= E\{y[n]y[n+1]\} = E\{(x[n] + q[n])(x[n+1] + q[n+1])\} = r_{xx}(1) \end{aligned}$$

Therefore, the measurement noise will affect only the autocorrelation values for k=0 (see answer a) above), and thus

$$r_{yy}(k) = \begin{cases} r_{xx}(k) + \sigma_q^2 = 2 & k = 0 \\ r_{xx}(k) & k > 0 \end{cases}$$

ii) Using Part a ii) above we have

$$\begin{bmatrix} r_{yy}(0) & r_{xx}(1) \\ r_{xx}(1) & r_{yy}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{xx}(1) \\ r_{xx}(2) \end{bmatrix} \Leftrightarrow \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

Since $r_{xx}(0) = 1 + r_{xx}(0)$ the so calculated coefficients are inaccurate (see also Problem 1.7 in your P/A sheets dealing with AR processes).
iii) The easiest way would be to shift the lag index in the autocorrelation functions in the Yule-Walker equations by one, and to calculate the coefficients from

$$r(k) = a_1 r(k-1) + a_2 r(k-2), \quad k > 0$$

3. [**bookwork and a new example**]
a) The curvature refers to the negative of the second derivative of the log-likelihood function and is used to measure the goodness of an estimator by quantifying how narrow the pdf is. When the PDF is viewed as a function of an unknown parameter it is termed the "likelihood function". The "sharpness" of the likelihood function determines the accuracy with which the unknown parameter may be estimated. For example

$$ln\ p(x[0]; A) = -ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x[0] - A)^2$$

then

$$\frac{\partial ln\ p(x[0]; A)}{\partial A} = \frac{1}{\sigma^2}(x[0] - A)$$

and the curvature

$$-\frac{\partial^2 ln\ p(x[0]; A)}{\partial A^2} = \frac{1}{\sigma^2}$$

Also the log-likelihood is advantageous as it makes the derivations more mathematically tractable, by replacing the exponential terms with linear terms, and products with sums. The logarithm is a monotonic operation and it keeps the relative positions of e.g. minima and maxima.
b) i) The system model for a vector parameter is given by

$$\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n], i = 1, \ldots, p \Rightarrow \hat{\boldsymbol{\theta}} = \mathbf{Ax}$$

and inherits the constrains from the scalar BLUE, that is unbiased and linear in the data, where the unbiased constraint

$$E\{\hat{\theta}_i\} = \sum_{n=0}^{N-1} a_{in} E\{x[n]\} = \theta_i \quad \Rightarrow \quad E\{\hat{\boldsymbol{\theta}}\} = \mathbf{A}E\{\mathbf{x}\} = \boldsymbol{\theta}$$

Recall that for every $\theta_i \in \boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$ we have

$$\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n], \quad i = 1, 2, \ldots, p \quad \text{and} \quad \mathrm{E}\{\hat{\theta}_\mathrm{i}\} = \sum_{\mathrm{n}=0}^{\mathrm{N}} \mathrm{a_n E\{x[n]\}} = \theta_\mathrm{i}$$

and the linear constraint stems from: $E\{x[n]\} = s[n]\theta \quad \Rightarrow \quad E\{\mathbf{x}\} = \mathbf{H}\boldsymbol{\theta}$ ↬ the constraint $\mathbf{AH} = \mathbf{I}$, where $\mathbf{A} = [a_{in}]_{(p \times N)}$ and $\mathbf{H}$ is a vector/matrix of terms $\{s[n]\}$
The vector BLUE now becomes

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^{-1} \mathbf{C}^{-1} \mathbf{x}$$

with the covariance matrix $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1}$
If the data are truly Gaussian, as in

$$\mathbf{x} = \mathbf{H}\hat{\boldsymbol{\theta}} + \mathbf{w} \quad \text{with} \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

then the BLUE also yields the Gauss-Markov theorem.
ii) The Gauss-Maarkov Theorem. Given

$$\mathbf{x} = \mathbf{H}\ \hat{\boldsymbol{\theta}} + \mathbf{w}$$

with $\mathbf{w}$ having zero mean and covariance $\mathbf{C}$, otherwise an arbitrary PDF, the vector BLUE can be found as:-

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\ \mathbf{C}^{-1}\ \mathbf{H}\right)^{-1}\ \mathbf{H}^{-1}\ \mathbf{C}^{-1}\, \mathbf{x}$$

and for every $\theta_i \in \boldsymbol{\theta}$, the minimum variance of θ_i is

$$var(\hat{\theta}_i) = \left[\left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1}\right]_{ii}$$

with covariance matrix of $\hat{\boldsymbol{\theta}}$

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T\, \mathbf{C}^{-1}\, \mathbf{H}\right)^{-1}$$

iii) These estimators would give the same solution for a DC level in WGN, and would attaind the CRLB. This is easily shown based on the linear model interpretation of the estimatino of vector processes.
c) The unknown parameter is the wind speed, whose estimator is given by

$$\hat{h} = \frac{1}{20} \sum_{i=1}^{20} \hat{h}_i$$

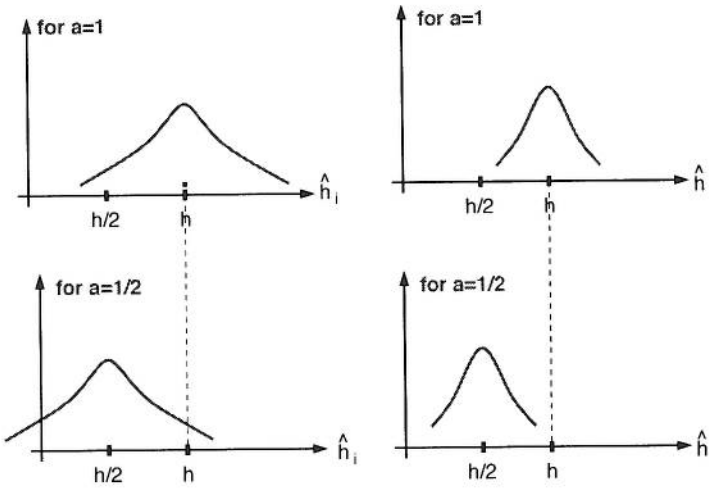

Figure 2: Estimated *pdf*'s for $a = 1$ and $a = 1/2$. Left hand graphs: original *pdf*'s. Right hand graphs: *pdf*'s after averaging.

The statistical expectation operator $E\{\cdot\}$ is linear and therefore the expected value of the averaging estimator becomes

$$E\{\hat{h}\} = \frac{1}{20}\sum_{i=1}^{20} E\{h_i\} = ah$$

and the variance

$$var(\hat{h}) = var(\hat{h}_i)/20 = \frac{\sigma_h^2}{20}$$

The estimator efficiency for the cases of $a = 1$ and $a = 1/2$ are illustrated in Figure 2.

Observe that for $a = 1/2$ the averaging causes the PDF to be more heavily concentrated about the wrong value of h, decreasing the probability that $\hat{h}$ is close to h. For $a = 1$ the estimator is not biased, however, averaging is beneficial (recall sample mean example), as illustrated on the bottom right graph in Figure 2.

4. [**bookwork and new examples**]
a) The aim is to minimize the cost function

$$J(\theta) = \sum_{n=0}^{N-1}(x(n) - s(n))^2$$

where $s(n)$ is a deterministic data model and the minimisation if performed over all the available data points.
The disadvantages mostly come from the assumptions of zero mean data, linearity, and the deterministic signal model.

- Problem with signal mean. If the noise is not zero mean, then the sample mean estimator actually models $x[n] = A + w[n] + w'[n]$

  $A + E\{w[n]\} + w'[n] \quad w[n] \sim \text{non} - \text{zero mean noise} \quad \text{w'[n]} \sim \text{zero mean noise}$

  The presence of non-zero mean noise $w[n]$ biases the LSE estimator $\nrightarrow$ LS approach assumes that the observed data are composed of a deterministic signal and zero mean noise.

- Nonlinear signal model, for instance $s[n] = \cos 2\pi f_0 n$, where the frequency f_0 is to be estimated. The LSE criterion

  $$J(f_0) = \sum_{n=0}^{N-1} (x[n] - \cos 2\pi f_0 n)^2$$

  is highly nonlinear in $f_0 \rightarrow$ closed form minimisation is impossible.

  - For $s[n] = A\cos 2\pi f_0 n$, if f_0 is known and A is unknown, then we can use the LS method, as A is linear in the data
  - When estimating both A and f_0,the error is quadratic in A and non-quadratic in $f_0 \rightsquigarrow$ minimize J wrt A for a given f_0, reducing to the minimization of J over f_0 only (separable least squares).

b) i) Sequential least squares are introduced to reduce the computational complexity of least squares methods and to calculate the estimates sequentially, as the data become available, rather than in a block manner as performed by the LS methods. They are the basic learning machines and operate on the basis of

$$\text{(new estimate)} = \text{(old estimate)} + \text{(correction)}$$

ii) Consider the problem of estimating the DC signal level in noise , for which we have obtained the LSE

$$\hat{A}[N-1] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

If we now observe the new sample $x[N]$, the enhanced estimate

$$\begin{aligned} \hat{A}[N] &= \frac{1}{N+1}\sum_{n=0}^{N} x[n] = \frac{1}{N+1}\Big(\sum_{n=0}^{N-1} x[n] + x[N]\Big) \\ \Rightarrow \quad \hat{A}[N] &= \frac{N}{N+1}\hat{A}[N-1] + \frac{1}{N+1}x[N] \end{aligned}$$

Clearly $\hat{A}[N]$ can be calculated from $\hat{A}[N-1]$ together with the new observation $x[N]$ The solution can be rewritten as

$$\begin{aligned} \hat{A}[N] &= \hat{A}[N-1] + \frac{1}{N+1}\left(x[N] - \hat{A}[N-1]\right) \\ \text{New estimate} &= \text{Old estimate} + \underbrace{\text{Gain} \times \text{Error}}_{\text{correction}} \end{aligned}$$

ii) The sequential expression for the minimum mean squared error becomes

$$\text{from} \quad J_{min}[N-1] = \sum_{n=0}^{N-1} \left(x[n] - \hat{A}[N-1]\right)^2$$

$$\text{we have} \quad J_{min}[N] = \sum_{n=0}^{N} \left(x[n] - \hat{A}[N]\right)^2$$

and can be writen in a more intuitive form as

$$J_{\min}[N] = J_{\min}[N-1] + \frac{N}{N+1}\left(x[N] - \hat{A}[N-1]\right)^2$$

Notice that this quantity increases with N due to the need to fit more points with the same number of parameters.

c) i) From the LS cost function, and for $\mathbf{a} = [a_1, \ldots, a_p]^T$ and $\mathbf{x}(n) = [x(n-1), \ldots, x(n-p)]^T$ we have

$$J = \sum_{n=0}^{N-1} (\hat{x}(n) - x(n))^2 = \sum_{n=0}^{N-1} (\hat{x}(n) - \mathbf{a}^T\mathbf{x}(n))^T(\hat{x}(n) - \mathbf{a}^T\mathbf{x}(n))$$

we need to find the gradient of this deterministic cost function, set it to zero and find the AR coefficients from there. Given that $\hat{x}(n)$ is a predicted value of the true $x(n)$, and thus

$$\frac{\partial \hat{x}(n)}{\partial a_i} = x(n-i) \quad \Rightarrow \quad \nabla_{\mathbf{a}}\hat{x}(n) = \frac{\partial \hat{x}(n)}{\partial \mathbf{a}} = \mathbf{x}(n)$$

$$\nabla_{\mathbf{a}} J = \sum_{n=0}^{N-1} \left(\hat{x}(n) - \mathbf{a}^T\mathbf{x}(n)\right)\mathbf{x}(n) = \mathbf{0}$$

giving (for $\mathbf{R} = E\{\mathbf{x}\mathbf{x}^T\}$ both the Yule-Walker solution and the Wiener filter.
ii) An adaptive filter connected in the System Identification configuration would estimate the AR coefficients recursively (see Fig. 3). For instance, for an AR(2) model we would have the situation in Figure 4. The cost function of an adaptive

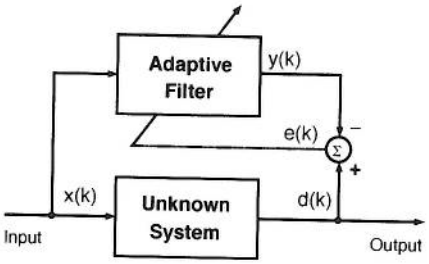

Figure 3: An adaptive filter in the system identification configuration

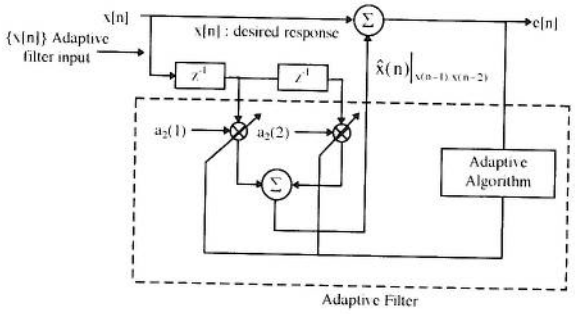

Figure 4: AR coefficients finding using an adaptive filter

filter is instantaneous $J = \frac{1}{2}e^2(n)$ as opposed to the deterministic cost function (sum of all errors squared) of the least squares method. The method of LS would therefore simply produce an estimate of the coefficients, which for white Gaussian w would give an unbiased minimum variance estimator. The estimate produced using an adaptive filter would be unbiased (if the filter is stable) whereas the weight estimates would fluctuate around the true weights, with the variance of the fluctuation proportional to the learning rate and filter length.

5) a) - Convergence in the mean (think of the requirement of an unbiased optimal weight estimate)

$$E\{\mathbf{w}(n)\} \rightarrow \mathbf{w}_0 \qquad as \quad n \rightarrow \infty$$

Convergence in the mean square (MS) (estimator variance, that is fluctuation of the instantaneous weight vector estimates around $\mathbf{w}_o$)

$$E\{e^2(n)\} \rightarrow \text{constant} \qquad \text{as} \quad n \rightarrow \infty$$

We can write this since the error is a function of the filter weights.

- The Mean Square (estimator variance) convergence condition is tighter: if LMS is convergent in the mean square, then it is convergent in the mean. The converse is not necessarily true. Convergence properties can also be inspected from the learning curves - a logarithmic plot of the mean squared error (MSE) along time, $10 \log e^2(n)$. In the sytem identification setting, learning curves can also be produced based on the misalignment $\| \mathbf{w}(n) - \mathbf{w}_o \|_2^2$

b) The cost (error,objective) function can be expanded as

$$\begin{aligned} J &= \frac{1}{2} E\{e\ e^T\} = \frac{1}{2} E\Big\{ \big(d - \mathbf{w}^T\mathbf{x}\big)\big(d - \mathbf{w}^T\mathbf{x}\big)^T \Big\} \\ &= \frac{1}{2} E\{d^2 - d\mathbf{x}^T\mathbf{w} - d\mathbf{w}^T\mathbf{x} + \mathbf{w}^T\mathbf{x}\mathbf{x}^T\mathbf{w}\} \\ &= \frac{1}{2} E\{d^2 - 2d\mathbf{x}^T\mathbf{w} + \mathbf{w}^T\mathbf{x}\mathbf{x}^T\mathbf{w}\} \\ &= \frac{1}{2} E\{d^2\} - \frac{1}{2}\, 2\mathbf{w}^T E\{\mathbf{x}d\} + \frac{1}{2}\, \mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} \end{aligned}$$

where the cross-correlation vector $\mathbf{p} \equiv E[\mathbf{x}d]^T$ and autocorr. matrix $\mathbf{R} \equiv E[\mathbf{x}\mathbf{x}^T]$
Thus, ($\mathbf{w}$ is a fixed vector) the cost function

$$J = \sigma_d^2 - 2\mathbf{w}^T\mathbf{p} + \mathbf{w}^T\mathbf{R}\mathbf{w}$$

is quadratic in $\mathbf{w}$ and for a full rank $\mathbf{R}$, it has one unique minimum.
Now: $\frac{\partial J}{\partial \mathbf{w}} = -\mathbf{p} + \mathbf{R}\cdot\mathbf{w} = \mathbf{0} \quad \Rightarrow \quad -\mathbf{p} + \mathbf{R}\cdot\mathbf{w}_o = \mathbf{0} \Rightarrow \quad \mathbf{w}_o = \mathbf{R}^{-1}\mathbf{p}$
(Wiener–Hopf Equation)

If the teaching signal is given by $d(n) = \mathbf{x}^T\mathbf{w}_o + q(n)$ then the minimum achievable mean squared error is σ_q^2, and this is also reflected in the cross-correlation $\mathbf{p}$ when calculating the Wiener filter above (simplifying the expression).

c) i) The DFT sequence of length N is given by

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk} = \sum_{n=0}^{N-1} x(n) W_N^{nk}$$

where $W_N = e^{-j\frac{2\pi}{N}}$. If we define

$$\mathbf{w_k^H} = \Big[\ 1, W_N^k, W_N^{2k}, \ldots, W_N^{k(N-1)}\ \Big]$$

where $X(k)$ is the inner product

$$X(k) = \mathbf{w_k^H}\cdot\mathbf{x} \tag{1}$$

Arranging the DFT coefficients in a vector we have

$$\mathbf{X} = \begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(N-1) \end{bmatrix} = \begin{bmatrix} \mathbf{w_0^H x} \\ \mathbf{w_1^H x} \\ \vdots \\ \mathbf{w_{N-1}^H x} \end{bmatrix} = \mathbf{W}\mathbf{x}$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{w_0^H} \\ \mathbf{w_1^H} \\ \vdots \\ \mathbf{w_{N-1}^H} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1\ldots & 1 \\ 1 & W_N & W_N^2\ldots & W_N^{N-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} \ldots & W_N^{(N-1)^2} \end{bmatrix}$$

ii) The matrix $\mathbf{W}$ is symmetric and nonsingular. In addition, due to orthogonality of the complex exponentials,

$$\mathbf{w_k^H} \cdot \mathbf{w_l} = \sum_{\mathbf{n=0}}^{\mathbf{N-1}} \mathbf{e}^{-\mathbf{j}\frac{2\pi}{\mathbf{N}}\mathbf{(k-l)}} = \begin{cases} N & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

it follows that $\mathbf{W}$ is orthogonal.
Due to orthogonality of $\mathbf{W}$, the inverse is

$$\mathbf{W^{-1}} = \frac{\mathbf{1}}{\mathbf{N}}\mathbf{W^H}$$

This explains the form for the inverse Discrete Fourier Transform (DFT)

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{j\frac{2\pi}{N}kn}$$

iii) For noisy data we are observing $x(n) + w(n)$ and the problem boils down to finding the solution that minimizes the mean squared error between the data model (DFT or iDFT) and the observed data. This can be performed by routine Wiener filtering. This can also be performed using an adaptive filter in a system identification setting, as it would identify the true Fourier coefficients. This filter would be complex valued.