E3.08
ISE3.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2005

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

ADVANCED SIGNAL PROCESSING

Monday, 25 April 10:00 am

Time allowed: 3:00 hours

Corrected Copy

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.
Second Marker(s) : R. Nabar

1) Consider the autoregressive moving average (ARMA) modelling.

a) State the expression for a general second order autoregressive $AR(2)$ process $z[n]$. [2]

i) Derive the autocorrelation function for this process. Can you write the expression for the autocorrelation function directly from the expression for the AR(2) process? Explain. [4]

ii) What is the set of stability conditions for this process (stability triangle)? Explain the bounds on the AR parameters for an AR(2) process to be stable. What are the four possibilities for a general shape of the autocorrelation function and spectrum? [6]

b) Consider the process given by

$$z[n] = 0.9z[n-1] - 0.4z[n-2] + w[n]$$

where $w[n]$ denotes samples of white Gaussian noise. Is the process $z[n]$ stable? [2]

In order to determine the AR parameters for process $z[n]$, state

i) the Yule-Walker equations. [3]

ii) the spectrum. Is the spectrum dominated by peaks, or it is flat? (Hint: we have an all–pole system) [3]

2) Consider the problem of least squares (LS) estimation.

a) State the least squares error criterion. What is the goal of least squares estimation with respect to the least squares error criterion? [4]

   i) For linear least squares explain how the dimensions of the signal and data spaces in LS estimation are related. [4]

   ii) Explain how the columns of the observation matrix $\mathbf{H}$ are related to the signal model and what makes it possible for the error to be orthogonal to the estimate. [3]

   iii) Wiener–Hopf and Yule–Walker are solutions of the least squares problem. Which one is related to the deterministic and which one to the stochastic error function J (in terms of some measure of the output error $e(n)$)? [2]

b) We want to model signal $x(n)$ using an all–pole model of the form

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_p(k) z^{-k}}$$

   i) Derive the normal equations that define the coefficients $a_p(k)$ that minimise the error (Hint: Found by setting the derivatives of E_p with respect to coefficients $a_p(i), \quad i = 1, \ldots, p$ to zero)

$$E_p = \sum_{n=0}^{\infty} |e(n)|^2$$

   where

$$e(n) = x(n) + \sum_{l=1}^{p} a_p(l) x(n-l)$$

   and derive the expression for the minimum error (Hint: use the orthogonality condition). [7]

3) State the aim of minimum variance unbiased (MVU) estimation. [2]

a) Give a simple example of a likelihood function for a random signal x. [2]

i) Explain in your own words the meaning of "sharpness" of this function. [2]

ii) How do we quantify the sharpness of a likelihood function? What is the relation between the Fisher information matrix and curvature of the log-likelihood function? [2]

iii) Define the "regularity condition" within the Cramer–Rao Lower Bound (CRLB) for scalar parameter. Give an interpretation of this condition. [2]

iv) Define an efficient estimator. [2]

b) Describe the need for linear models within MVU estimation. Write down the expression for a simple first order random linear model. [2]

i) Write down the expression for a linear model in a compact vector-matrix notation and explain the terms in that equation. [2]

ii) State the Cramer Rao Lower Bound theorem for linear models. [2]

iii) A linear model is used for system identification and is given by

$$x[n] = \sum_{k=0}^{p-1} h[k]u[n-k] + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $x[n]$ is the output of the system, $h[n]$ are the unknown filter coefficients, $u[n]$ is filter input, and $w[n]$ denotes additive white Gaussian noise. Write down this data model in its vector–matrix form. Denote the observation matrix and the parameter vector. [2]

4) State the problem of optimum linear filtering (Wiener problem). [2]

a) Derive the method of steepest descent. [2]

i) Define and sketch a simple error surface. [2]

ii) Explain the role of learning rate. [2]

iii) What are the advantages of this method over the standard Wiener filter? [2]

iv) Write down the cost function for the steepest descent method. [2]

b) The coefficient update equation of least mean square (LMS) adaptive filter is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n)\mathbf{x}(n)$$

where $\mathbf{w}$ is the coefficient vector, μ is the learning rate, $\mathbf{e}$ is the filter error and $\mathbf{x}$ is the input vector. The block LMS algorithm, on the other hand, accumulates these corrections for L samples, beginning at time n while holding the weight vector $\mathbf{w}$ constant. A correction term is then applied at the end of the block to form an update at time $(n+L)$ as follows

$$\mathbf{w}(n+L) = \mathbf{w}(n) + \mu \sum_{l=0}^{L-1} e(n+l)\mathbf{x}(n+l)$$

where

$$e(n+l) = d(n+l) - \mathbf{w}^T(n)\mathbf{x}(n+l), \quad l = 0, 1, \ldots, L-1$$

i) In your own words, discuss the advantages/disadvantages of the block LMS algorithm compared to the standard LMS. [4]

ii) By evaluating the behaviour of $E\{\mathbf{w}(n)\}$ as a function of n, discuss the conditions for the step size μ so that the filter converges in the mean ($E\{\mathbf{w}(n)\} \rightarrow$ constant, when $n \rightarrow \infty$). [4]

5) A simple extension of linear finite impulse response (FIR) adaptive filters is a nonlinear FIR filter shown in the Figure 5.1. The nonlinearity Φ is a saturation–type nonlinear function, such as tanh, and the output of this filter is given by $y(k) = \Phi(\mathbf{x}^T(k)\mathbf{w}(k))$. In the stochastic gradient setting, the cost function for this filter is based on the minimisation of the squared instantanous output error and is given by

$$E(k) = \frac{1}{2}e^2(k)$$

a) Give the reasons for this nonlinear FIR filter also being called a "dynamical neuron". [4]

b) Derive the weight update equation for this filter based on the cost function given above and the gradient descent approach (Hint: $\Delta\mathbf{w}(k) = -\eta\nabla_{\mathbf{w}}E(k)$). [8]

c) Explain the difference in the way this filter and the standard FIR adaptive filter trained by the LMS algorithm process signals with large dynamical ranges. Which structure do you expect to perform better when filtering nonlinear signals? [2]

d) If the nonlinear function Φ is the tanh function, explain the effect of the saturation within the output of the filter has on the learning process (Hint: Cases when the operating point moves towards the tails of the nonlinearity, where the gradient values are very small). [2]

i) What is the effect of saturation–type nonlinearity on the output magnitude range? [2]

ii) Propose a way to improve the output range of this structure. [2]

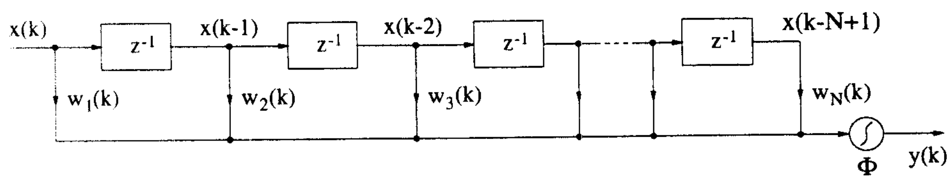

Figure 5.1: A nonlinear FIR filter

Solutions:

1) a)

$$z[n] = a_1 z[n-1] + a_2 z[n-2] + w[n]$$ [2]

where a_1, a_2 are the model parameters and $\{w[n]\}$ is the driving white noise.

i) By applying the expectation operator $E\{\cdot\}$ to

$$z[n-k]z[n]$$ [4]

we see that the ACF model follows the general form of the AR(2) process. Therefore we have

$$\rho(k) = a_1\rho(k-1) + a_2\rho(k-2)$$

where $\rho(0) = 1$.

ii) Using the results for AR(1) and extending for AR(2) model, we can derive the bounds on stability for AR(2) processes, which in a convenient way can be put within a "stability triangle" shown below. Obviously the stability conditions are $-2 \leq a_1 \leq 2, -1 \leq a_1 \leq 1$. [6]

b) The process is stable, by inspection from the above stability triangle.
i) The second order variant of the general Yule Walker solution.
Substituting $p = 2$ into the general form of Yule–Walker equations, we have [2]

$$\rho_1 = a_1 + a_2\rho_1$$ [3]
$$\rho_2 = a_1\rho_1 + a_2$$

which when solved for a_1 and a_2 gives

$$a_1 = \frac{\rho_1(1-\rho_2)}{1-\rho_1^2}$$
$$a_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2}$$

ii)

$$P_{zz}(f) = \frac{2\sigma_w^2}{|1 - a_1 e^{-\jmath 2\pi f} - a_2 e^{-\jmath 4\pi f}|^2}$$

$$= \frac{2\sigma_w^2}{1 + a_1^2 + a_2^2 - 2a_1(1 - a_2\cos(2\pi f) - 2a_2\cos(4\pi f))}, \qquad 0 \leq f \leq 1/2$$ [3]

Due to the all–pole system, the spectrum is dominated by peaks.

2/9

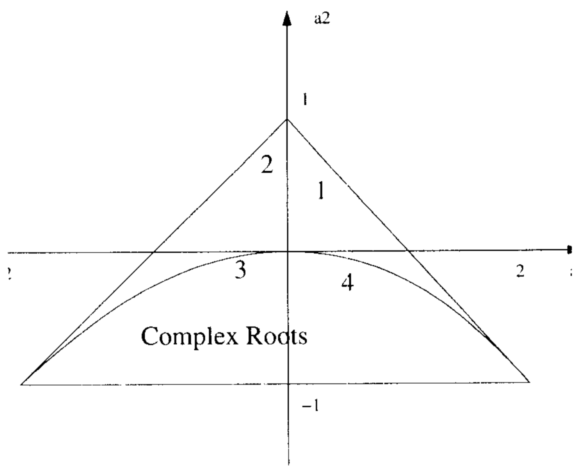

Figure 1: AR(2) Stability triangle. Region 1: Decaying ACF, Region 2: Decaying oscillating ACF, Region 3: Oscilating pseudoperiodic ACF, Region 4: Pseudoperiodic ACF

2) a)

- The Least Squares Estimator of θ chooses the value that makes $s[n]$ closest to the observed data $x[n]$, where closeness is measured by the LS error criterion [4]

$$J(\theta) = \sum_{n=0}^{N-1} \underbrace{(x[n] - s[n])^2}_{e[n]}$$

LSE: $\min_\theta$ J(θ)

Note. no probabilistic assumptions have been made about the data $x[n]$ i) The signal sub–space is spanned over the columns of the observation matrix and is normally of lower dimension than the data space. [4]

ii) The columns of the observation matrix define the vector space in which the signal resides. Therefore signal model is built as a linear combination of those vectors. Therefore the estimated values are a projection onto the signal subspace, and the estimation error is therefore orthogonal to the estimated signals. [3]

iii) We draw a distinction between stochastic and deterministic measures [2]

(a) Stochastic

$$J = \min_{\mathbf{h}} E\{|e(n)|^p\}$$

(b) Deterministic

$$J = \min_{\mathbf{h}} \sum_{n} |e(n)|^p$$

Wiener–Hopf – stochastic, Yule–Walker – deterministic.
b) i)

$$\frac{\partial E_p}{\partial a_p(k)} = \sum_{n=0}^{\infty} 2e(n)x(n-k) = 0$$

[7]

$$\sum a_p(l) \left[\sum_{n=0}^{\infty} x(n-l)x(n-k) \right] = -\sum_{n=0}^{\infty} x(n)x(n-k)$$

Define

$$r_x(k,l) = \sum_{n=0}^{\infty} x(n)x(n-k)$$

Thus the normal equations become

$$\sum_{l=1}^{p} a_p(l) r_x(k-l) = -r_x(k)$$

The minimum error becomes (using the orthogonality condition)

$$E_{pMIN} = r_x(0) + \sum_{l=1}^{p} a_p(l) r_x(l)$$

4 | 9

3) To formulate lower bound on the variance of any unbiased estimator. [2]

a) The likelihood function is a probability density function defined in terms of the unknown parameter to be estimated, that is $p(x, \theta)$. For example, for a single sample $x[0] = A + w[0]$, where $w[0] \sim N(0, \sigma^2)$. [2]

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

i) Bu sharpness we mean the variance of the likelihood function. The "narrower" the likelihood function the better the estimate. [2]

ii) Via the so called "curvature". This is the negative of the second partial derivative of the log likelihood function at its paek, with respect to the unknown parameter, given by [2]

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

Therefore, as expected, the curvature increases as σ^2 decreases. The Fisher information is related to the expected value of curvature. [2]

iii) This is the necessary condition within the CRLB theorem.
The PDF $p(\underline{x}; \theta)$ satisfies the "regularity" condition if

$$E\left[\frac{\partial ln\ p(x[0]; \theta)}{\partial \theta}\right] = 0 \qquad \forall \theta$$

where the expectation is taken with respect to $p(\underline{x}; \theta)$. THis is closely related to the bias of the estimator. [2]

iv) An estimator which is unbiased and attains the CRLB is said to be "efficient" in that it attains the CRLB and efficiently uses that data.

b) [2]

- Generally it is difficult to determine the MVU estimator
- In signal processing, however, a linear data model can often be employed for which it is straightforward to determine the MVU estimator
- Simple example is to fit a straight line through noise corrupted data; assume

  $$x[n] = A + Bn + w[n] \qquad n = 0, 1, \ldots, N-1$$

  where $w[n] \sim N(0, \sigma^2)$ B - slope and A - intercept

i) This data model can be written more compactly in matrix notation as [2]

$$\underline{x} = H\underline{\theta} + \underline{w}$$

5|9

where $\underline{x} = [x[0]\, x[1] \ldots x[N-1]]^T$

$$H = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix}$$

$$\underline{\theta} = [AB]^T$$
$$\underline{w} = [w[0]\, w[1] \ldots w[N-1]]^T$$
$$\underline{w} \sim N(\underline{0}, \sigma^2 \underline{I})$$

[2]

ii) If the observed data can be modelled as

$$\underline{x} = H\underline{\theta} + \underline{w}$$

where
- $\underline{x}$ is "observation factor"
- H is N×p "observation matrix" (N×p) Rank p
- $\underline{\theta}$ is p×1 "parameter vector"
- $\underline{w}$ is N×1 "noise vector" $\sim$ N$(\underline{0}, \sigma^2 \underline{I})$

The MVU Estimator is

$$\underline{\theta} = \left(H^T H\right)^{-1} H^T \underline{x}$$

with covariance matrix

$$C_{\hat{\underline{\theta}}} = \sigma^2 \left(H^T H\right)^{-1}$$

Note the statistical performance of $\hat{\underline{\theta}}$ is completely satisfied because $\hat{\underline{\theta}}$ is a linear transformation of a Gaussian vector $\underline{x}$, i.e.

$$\hat{\underline{\theta}} \sim N\left(\underline{\theta}, \sigma^2 \left(H^T H\right)^{-1}\right)$$

iii) [2]

$$\underline{x} = \underbrace{\begin{bmatrix} u[0] & 0 & \ldots & 0 \\ u[1] & u[0] & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \ldots & u[N-p] \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(p-1) \end{bmatrix}}_{\underline{\theta}} + \underline{w}$$

$$\underline{w} \sim N(\underline{0}, \sigma^2 \underline{I})$$

6|9

4) Determine the optimum set of weights for which the cost function attains its minimum. [2]

a) i), ii), iii) and iv) [2+2+2+2]

Problem with Wiener filter: Matrix inversion of the $p \times p$ matrix $\mathbf{R}$.

We may avoid the need for matrix inversion by using the method of steepest descent.

Error surface: a quadratic surface defined by the cost function, for varying weight vectors. The weights here assume a **time–varying** form, and their values are adjusted in an **iterative** fashion allong the error surface with the aim of moving them progerssively toward the optimum solution.

It is intuitively reasonable that successive adjustments applied to the tap weights of the filter be in the direction of steepest descent of error surface, that is in a *direction opposite to the gradient vector* whose elements are defined by $\nabla_{w_k} J$, $k = 1, 2, \ldots, p$ In a corresponding way, the gradient of the error sur-

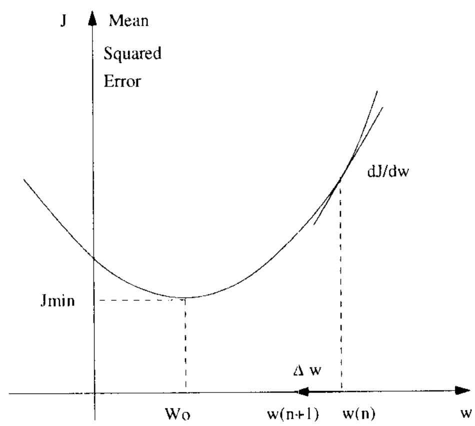

Figure 2: The method of steepest descent

7 | 9

face of the filter wrt the weights takes on a *time varying* form

$$\nabla_{w_k} J(n) = -r_{dx}(k) + \sum_{j=1}^{p} w_j(n) r_x(j,k)$$

(indices j, k refer to locations of different sensors in space, whereas the index n refers to iteration number). According to the method of steepest descent, tje adjustment applied to the weight $w_k(n)$ at iteration n is defined by

$$\Delta w_k(n) = -\eta \nabla_{w_k} J(n), \qquad k = 1, 2, \ldots, p$$

where η is a positive constant called the **learning rate** parameter (step size). Given the **old** value of the kth weight $w_k(n)$ at iteration n, the **updated** value of this weight at the next iteration $(n+1)$ is computed as

$$w_k(n+1) = w_k(n) + \Delta w_k(n) = w_k(n) - \eta \nabla_{w_k} J(n)$$

or

The updated value of the Wiener filter = the old value + correction

Finally, we have

$$w_k(n+1) = w_k(n) + \eta \left[r_{dx}(k) - \sum_{j=1}^{p} w_j(n) r_x(j,k) \right], \qquad k = 1, \ldots, p$$

The SD method is **exact** in the sense that there are no approximations made in the derivation.
The derivation is based on minimizing the mean squared error

$$J(n) = \frac{1}{2} E\{e^2(n)\}$$

The cost function is an *ensemble average* taken at time n over an ensemble of *spatial filters*. The SD method can also be derived by minimizing the *sum of error squares*

$$\mathcal{E}_{total} = \sum_{i=1}^{n} \mathcal{E}(i) = \frac{1}{2} \sum_{i=1}^{n} e^2(i)$$

b) i) Advantage: more accurate than LMS (due to averaging), and lower misadjustment. Disadvantage: more difficult to track rapidly varying processes. [4]
ii)

$$E\{\mathbf{w}(n+L)\} = E\{\mathbf{w}(n)\} + \mu \sum_{l=0}^{L-1} E\{e(n+l)\mathbf{x}(n+l)\}$$ [4]

$$= E\{\mathbf{w}(n)\} + \mu \sum_{l=0}^{L-1} E\{[d(n+l) - \mathbf{w}^T(n)\mathbf{x}(n+l)]\mathbf{x}(n+l)\}$$

$$= (\mathbf{I} - \mu L \mathbf{R}_{xx}) E\{\mathbf{w}(n)\} + \mu L \mathbf{r}_{dx}$$

8/9

Therefore, we set first term to zero to obtain

$$0 < \mu < \frac{2}{L\lambda_{max}}$$

It suffices to state the above expectation and describe in your own words that we need to minimise it.

9|9

5) a) Or Dynamical perceptron. The structure us an electrical model of a neuron from the brain. It has its synaptic part (delayed inputs and weihts) and somatic part (summation and nonlinearity). [4]

[8]

b) Identical procedure as for the LMS, with the exception that we need to account for the nonlinearity within the structure. The final update is

E

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \eta e(k)\Phi'(\mathbf{x}^T(k)\mathbf{w}(k))\mathbf{x}(k)$$

[2]

c) The FIR filter trained by LMS is linear and is not sensitive to the amplitudes of the signal. This filter has a saturation type nonlinearity and is sensitive to the changes in the signal dynamics. The nonlinearity can be thought as having a quasilinear range and a saturation range, hence nonlinear distortion of signals. [2+2+2]
i) The output magnitude range is limited to the range of the nonlinearity
ii) Have an adaptive amplitude activaton function, or other forms of range reduction prior to filtering and range extension after filtering.