final

E3.08
ISE3.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2010

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

ADVANCED SIGNAL PROCESSING

Thursday, 6 May 10:00 am

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.L. Dragotti, P.L. Dragotti

1) Consider the problem of estimating the value of a parameter, θ, from a sequence of random variables $x[n], \quad n = 1, 2, \ldots, N$. Since the estimate is a function of N random variables, we will denote it by $\hat{\theta}_N$.

a) Define the bias B in parameter estimation. When do we say that an estimate is unbiased? [2]

b) Define an asymptotically unbiased estimator. The data $\{x[0], \ldots, x[N-1]\}$ are independent and identically distributed (IID) as $\mathcal{N}(0, \sigma^2)$. We wish to estimate the variance σ^2 as

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]$$

Is this an unbiased estimator? [2]

c) Define the terms *mean square convergence* and *consistent estimator.* [2]

i) Is the sample mean estimate $\hat{m}_x = \frac{1}{N} \sum_{n=1}^{N} x[n]$ unbiased and consistent? [2]

ii) Is the estimator in part b) consistent? [2]

d) Let x be the random variable defined on the coin flipping experiment, with $x = 1$ if the outcome is heads and $x = -1$ if the outcome is tails. The coin is unfair so that the probability of flipping *heads* is p and the probability of flipping *tails* is $(1 - p)$.

i) Find the mean of x. [2]

ii) Suppose the value for p is unknown and that the mean of x is to be estimated. Flipping the coin N times and denoting the resulting values for x by $x[i], \quad i = 1, \ldots, N$, consider the following estimator for m_x

$$\hat{m}_x = x[N]$$

Is this estimator unbiased? [4]

iii) Find the variance of the estimator from part ii). Is this estimator consistent? [4]

2) Consider the problem of mixed autoregressive moving average (ARMA) modelling.

a) For a general ARMA(p, q) process:

i) Derive the expression for the autocorrelation function $r_{xx}[k]$ of this process and find the expression for the autocorrelation function $r_{xx}[k]$ for $k \geq q+1$. [6]

ii) State and explain the equation for the power spectrum of a general ARMA(p, q) process. [4]

b) We desire to use the AR(1) model

$$x(n) = ax(n-1) + w(n), \qquad w(n) \sim \mathcal{N}(0, \sigma_w^2)$$

to predict the process $x(n)$ based on the previous sample $x(n-1)$, that is

$$\hat{x}(n) = ax(n-1)$$

Using the orthogonality principle (error is orthogonal to the data) or otherwise, find the optimal value of the parameter a and the minimum prediction error power. [6]

c) Prove that the forward and backward AR processes

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + w[n]$$

and

$$x[n] = \sum_{k=1}^{p} a_k x[n+k] + w[n]$$

where $w[n] \sim \mathcal{N}(0, \sigma_w^2)$ have the same Power Spectral Densities (PSD)s. [4]

3) Consider the problem of Maximum Likelihood Estimation (MLE).

a) Explain the principle of MLE. [4]

i) What are the properties of MLE if an efficient estimator does not exist? [4]

ii) Illustrate the operation of MLE on the example of estimating a DC level in white Gaussian noise. [4]

b) Assuming a scalar parameter, if an efficient estimator exists, then we have

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = I(\theta)\big(\hat{\theta} - \theta\big)$$

where $\mathbf{x}$ is the data vector, I is the Fisher information, θ is an unknown parameter, and $\hat{\theta}$ is the estimated value of θ. Show that the maximum likelihood method will produce this efficient estimator.
Hint: compare with the regularity condition of the Cramér-Rao lower bound. [2]

c) Consider the real linear model for $n = 0, 1, \ldots, N-1$, given by

$$x[n] = \alpha + \beta n + z[n], \qquad z \sim \mathcal{N}(0, \sigma_z^2)$$

i) Evaluate the Maximum Likelihood Estimate (MLE) of the unknown parameters - slope β and the intercept α, by assuming the Gaussian $z[n]$ with zero mean and variance σ_z^2. [4]

$$\text{Hint:} \qquad p(\mathbf{x};\alpha;\beta) = \frac{1}{(2\pi)^{N/2}\sigma_z^N}\, e^{\frac{-1}{2\sigma_z^2}\sum_{n=0}^{N-1}(x[n]-[\alpha+\beta n])^2}$$

ii) Find the MLE of α if we set $\beta = 0$. [2]

4) Consider the method of least squares (LS).

a) State the least squares optimisation problem for the estimation of a vector parameter. [4]

b) Derive the LS solution for the vector parameter and explain the role of the observation matrix $\mathbf{H}$. In your own words comment on the physical meaning of the vectors making up the columns of the observation matrix. [6]

c) We would like to build a predictor of digital waveforms. Such a system forms an estimate of a later sample (say n_0 samples later) by observing p consecutive data samples, and is given by

$$\hat{x}[n+n_0] = \sum_{k=1}^{p} a_p[k]x[n-k]$$

The predictor coefficients $a_p[k]$ are to be chosen to minimize

$$E_p = \sum_{n=0}^{\infty} \left(x[n+n_0] - \hat{x}[n+n_0]\right)^2$$

Derive the equations that define the optimum set of coefficients $a_p[k]$. [6]

d) Discuss the advantages of using the method of least squares. [4]

5) Consider a linear finite impulse response adaptive filter.

a) Derive the method of steepest descent for the adaptation of filter weights. [8]

b) Draw a block diagram and explain the operation of the adaptive prediction configuration. [4]

c) Suppose that the input to an adaptive linear predictor is white noise with an autocorrelation sequence $r_x(k) = \sigma_x^2 \delta(k)$, where the symbol $\delta(k)$ is the delta function.

i) Derive the Wiener solution and find the p-th order optimal predictor of this process. What is the value of the crosscorrelation $\mathbf{r}_{dx} = E[d(n)\mathbf{x}(n)]$ of the teaching signal $d(k)$ and the input vector $\mathbf{x}(n)$ in this prediction configuration, and for the white input used? [2]

ii) Describe the evolution of the weight vector using the vector form of the method of steepest descent

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu\big[\mathbf{R}_x \mathbf{w}(n) - \mathbf{r}_{dx}\big]$$

with the stepsize $\mu = 1/(5\sigma_x^2)$ and the initial weigh vector $\mathbf{w}_0 = [1, 1, \ldots, 1]^T$. Does the steepest descent method converge to the solution found in part i)? [4]

Hint: the autocorrelation matrix is $\mathbf{R}_x = \sigma_x^2 \mathbf{I}$.

d) Based on the standard cost function $J(n) = E[e^2(n)]$, one variant of Newton's algorithm is

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n) \mathbf{R}_x^{-1} \mathbf{x}(n)$$

Compare with the update of the Least Mean Square (LMS) algorithm and discuss the direction of the gradient on the error surface. [2]

E3.08
ISE3.17

Advanced Signal Processing - 2010

1
11

Solutions

1) [**Bookwork and practical application of bookwork**]

1) a) Bias $B = \theta - E\{\hat{\theta}_N\}$.
If the bias is zero, then the expected value of the estimate is equal to the true value, i.e. $E\{\hat{\theta}_N\} = \theta$ and the estimator is said to be unbiased.

b) [**bookwork and new example**]
An estimator is asymptotically unbiased if an estimate is biased but the bias goes to zero as the number of observations, N goes to infinity, that is

$$\lim_{N\to\infty} E\{\hat{\theta}_N\} = \theta$$

For the estimator of the variance, we have

$$E\{\sigma^2\} = E\Big\{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\Big\} = \frac{1}{N}\sum_{n=0}^{N-1} E\{x^2[n]\} = \sigma^2$$

Therefore, this estimator is unbiased.

c) [**bookwork and new example**]
An estimate $\hat{\theta}_N$ is said to converge to θ in the mean square sense if

$$\lim_{N\to\infty} E\{|\hat{\theta}_N - \theta|^2\} = 0$$

An estimate is consistent if it is unbiased and it converges in the mean square sense.

i) The sample mean estimate is unbiased, since $E\{\hat{m}_x\} = \frac{1}{N}\sum_{n=1}^{N} E\{x[n]\} = m_x$.
Since the variance of the sample mean estimate is

$$var\{\hat{m}_x\} = \frac{1}{N^2}\sum_{n=1}^{N} var\{x[n]\} = \frac{\sigma_x^2}{N}$$

which goes to zero as $N \to \infty$, it follows that the sample mean is a consistent estimator.

ii) Since the data are IID, from

$$\mathrm{var}\{\hat{\sigma}^2\} = \frac{1}{\mathrm{N}^2}\mathrm{var}\Big\{\sum_{\mathrm{n}} \mathrm{x}^2[\mathrm{n}]\Big\} = \frac{1}{\mathrm{N}^2}\mathrm{N}\mathrm{var}\big\{\mathrm{x}^2[\mathrm{n}]\big\} = \frac{1}{\mathrm{N}}\mathrm{var}\big\{\mathrm{x}^2[\mathrm{n}]\big\}$$

which goes to zero as $N \to \infty$, indicating a consistent estimator.
The variance of $x^2[n]$ can be calculated as

$$\mathrm{var}\{\mathrm{x}^2[\mathrm{n}]\} = \mathrm{E}\{\mathrm{x}^4[\mathrm{n}]\} - \mathrm{E}\{\mathrm{x}^2[\mathrm{n}]\}^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4 \quad \Rightarrow \quad \mathrm{var}\{\hat{\sigma}^2\} = \frac{2\sigma^4}{\mathrm{N}} \to 0 \ \text{ for } \ \mathrm{N} \to \infty$$

d) [**new example**]
i) $m_x = E\{x\} = p \cdot 1 + (-1) \cdot (1-p) = 2p - 1$

ii) For an estimator $\hat{m}_x = x[N]$, the mean is

$$E\{\hat{m}_x\} = E\{x[N]\} = 2p - 1$$

and the estimator is unbiased. However, $\hat{m}_x = x[N]$ is not a good estimator of the mean.

iii) The estimate of the mean, $\hat{m}_x$ will either be equal to one, with probability p or it will be equal to minus one, with a probability of $(1-p)$. Therefore the accuracy of the estimate $E\{\hat{m}_x\} = x[N]$ does not improve as the number of observations N increases. The variance of the estimate

$$var\{\hat{m}_x\} = var\{x[N]\} = 4p(1-p)$$

does not decrease with N. The estimator does not converge in the mean square sense and is therefore not consistent.

2) a) [**bookwork and new examples**]
a) For an ARMA(p,q) random processes $x[n]$, the process $x[n]$ and driving noise $w[n]$ are related by a linear constant coefficient equation

$$x[n] = \sum_{l=1}^{p} a_p(l)x[n-l] + \sum_{l=0}^{q} b_q(l)w[n-l]$$

i) [**combination of bookwork and worked example**]
The autocorrelation function of $x[n]$ and crosscorrelation between $x[n]$ and $w[n]$ follow the same functional expression as that of an ARMA(p,q) model above. Multiply both sides of the above equation by $x[n-k]$ and apply the statistical expectation operator, to yield

$$r_{xx}[k] = \sum_{l=1}^{p} a_p(l)r_{xx}[k-l] + \sum_{l=0}^{q} b_q(l)r_{xw}[k-l]$$

Since for $k \geq q+1$, there is no correlation between $x[n]$ and $w[n]$, the ACF follows the AR part of the above equation, that is

$$r_{xx}(k) = \sum_{l=1}^{p} a_p(l)r_{xx}[k-l] \quad for \quad k \geq q$$

ii) [**combination of bookwork and worked example**]
From the $\mathcal{Z}$–domain representation of an ARMA(p,q) process, we have

$$H(z) = \frac{B_q(z)}{A_p(z)} = \frac{\sum_{k=0}^{q} b_q(k) z^{-k}}{1 + \sum_{k=1}^{p} a_p(k) z^{-k}}$$

Assuming that the filter is stable, the output process $x[n]$ will be wide–sense stationary and with $P_w = \sigma_w^2$, the power spectrum of $x[n]$ will be

$$P_x(z) = \sigma_w^2 \frac{B_q(z)B_q(z^{-1})}{A_p(z)A_p(z^{-1})}$$

or in terms of frequency θ

$$P_z(e^{\jmath\theta}) = \sigma_w^2 \frac{\left|B_q(e^{\jmath\theta})\right|^2}{\left|A_p(e^{\jmath\theta})\right|^2}$$

b) [**new example**]
Orthogonality Principle:- $(x[n] - \hat{x}[n]) \perp x[n-1]$

$$\begin{aligned} \Rightarrow E\{(x[n] - \hat{x}[n])x[n-1]\} &= 0 \\ E\{(x[n] - ax[n-1])x[n-1]\} &= 0 \\ \Rightarrow r_{xx}(1) &= ar_{xx}(0) \\ \Rightarrow a &= r_{xx}(1)/r_{xx}(0) \end{aligned}$$

$$\begin{aligned} \sigma_w^2 &= r_{xx}[0] - ar_{xx}[1] \qquad \text{Yule Walker Equation} \\ &= r_{xx}[0](1 - [\frac{r_{xx}[1]}{r_{xx}[0]}]^2) \end{aligned} \tag{1}$$

c) [**new example**]
Forward:

$$x[n] - \sum_{k=1}^{p} a_k x[n-k] = w[n]$$

$$\text{Z transform:} \qquad X(z)[1 - a_1 z^{-1} - \cdots - a_p z^{-p}] = W(z)$$

Backward:

$$x[n] - \sum_{k=1}^{p} a_k x[n+k] = w[n]$$

$$\text{Z transform:} \qquad X(z)[1 - a_1 z^{1} - \cdots - a_p z^{p}] = W(z)$$

The power spectrum of an AR process is given by

$$\begin{aligned} P_{xx}(z) &= \frac{\sigma_w^2}{A(z)A^*(\frac{1}{z^*})} \\ &= \frac{\sigma_w^2}{[1-a_1z^{-1}-\cdots-a_pz^{-p}][1-a_1z^1-\cdots-a_pz^p]} = W(z) \end{aligned}$$

and is independent of the forward or backward AR process.

3) a) [**bookwork**]

Principle: Estimate a parameter such that for this value the probability of obtaining an actually observed sample is as large as possible. This probability depends on a parameter which is adjusted to give it a maximum possible value.

- Let a random variable x have a probability distribution dependent on a parameter θ
- The parameter θ lies in a space of all possible parameters θ
- Let $p_x(x|\theta)$ be the probability density function of x given θ
- Assume that the mathematical form of p_x is known but not θ

The joint pdf of m sample random variables evaluated at each sample point $x_1, x_2, \ldots, x_m$ is given as

$$l(\theta, x_1, x_2, \ldots, x_m) = l(\theta, \mathbf{x}) = \prod_{i=1}^{m} p_x(x_i|\theta)$$

The above is known as the likelihood of the sampled observation. The Maximum Likelihood Principle requires us to select that value of θ that maximises the likelihood function.

It is often more convenient to use

$$\Lambda(\boldsymbol{\theta}, \mathbf{x}) = \log(p_x(\mathbf{x}|\boldsymbol{\theta})$$

The maximum is then at

$$\frac{\partial\Lambda(\boldsymbol{\theta}, \mathbf{x})}{\partial\boldsymbol{\theta}} = 0$$

i) [**bookwork**]

- If an efficient estimator exists the maximum likelihood procedure will produce it

- When an efficient estimator does not exist, the MLE has the desirable feature that it yields "an asymptotically efficient" estimator, that for sufficiently large datasets, is

  - unbiased
  - achieves the CRLB
  - has a Gaussian PDF, $\hat{\theta}^{asy} \sim \mathcal{N}(\theta, I^{-1}(\theta))$

  Provided the PDF $p(\underline{\boldsymbol{x}}; \theta)$ satisfies the regularity conditions:-

  - the derivatives of the log-likelihood function exist
  - and the Fisher information is non-zero

ii) [**worked example**]
D.C. level in WGN

$$x[n] = A + w[n], \qquad n = 0, 1, \ldots, N-1, \qquad w[n] \sim \mathcal{N}(0, \sigma^2)$$

PDF $p(\underline{\boldsymbol{x}}; A) = \dfrac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\dfrac{1}{2\sigma^2}\displaystyle\sum_{n=0}^{N-1}(x[n]-A)^2\right]$
Take the derivative of the log-likelihood function

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

Set the result to zero to yield the MLE

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

$\Rightarrow$ clearly the MVU estimator yields the CRLB (efficient).

b) [**new example**]
To find the MLE we maximise p or equivalently lnp, that is $\frac{\partial lnp}{\partial \theta} = 0$. But

$$\frac{\partial lnp}{\partial \theta} = I(\theta)\big(\hat{\theta} - \theta\big)$$

Since $I(\theta) > 0$ for all θ, and we always have to assume this - otherwise pdf does not depend on θ, the only solution is $\theta = \hat{\theta}$. Therefore MLE is just $\hat{\theta}$, the efficient estimator.

c) [**new example**]

$$z[n] = x[n] - (\alpha + \beta n) \sim \mathcal{N}(0, \sigma_z^2)$$

$$p(\mathbf{x};\alpha;\beta) = \frac{1}{(2\pi)^{N/2}\sigma_z^N} exp(\frac{-1}{2\sigma_z^2}\sum_{n=0}^{N-1}(x[n]-[\alpha+\beta n])^2)$$

Minimize $p(\mathbf{x};\alpha;\beta)$ with respect to α and β and equate to zero to yield

$$1/N\sum_{i=0}^{N-1}x[n] = \alpha + \frac{\beta}{N}\sum_{i=0}^{N-1}n = \alpha + \frac{\beta}{2}(N-1)$$

Similarly

$$\begin{aligned}\sum_{i=0}^{N-1}nx[n] &= \alpha\sum_{i=0}^{N-1}n + \beta\sum_{i=0}^{N-1}n^2 \\ \Rightarrow \frac{1}{N}\sum_{i=0}^{N-1}nx[n] &= \frac{\alpha}{2}(N-1) + \frac{\beta}{6}(N^2+N+1/2)\end{aligned}$$

Thus,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 1 & (N-1)/2 \\ N-1 & (N^2+N+1/2)/6 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_{i=\phi}^{N-1} x[n] \\ \sum_{i=\phi}^{N-1} nx[n] \end{pmatrix}$$

when $\beta = 0$, then

$$\hat{\alpha} = \frac{1}{N}\sum_{i=0}^{N-1}x[n] \qquad \text{(the sample mean)}$$

4) a) [**bookwork**] LSE is found by minimising

$$\begin{aligned}J(\boldsymbol{\theta}) &= \sum_{n=0}^{N-1}\left(x[n]-s[n]\right)^2 = \left(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\right)^T\left(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}\right) \\ &= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}]\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} \\ \frac{\partial J(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}\end{aligned}$$

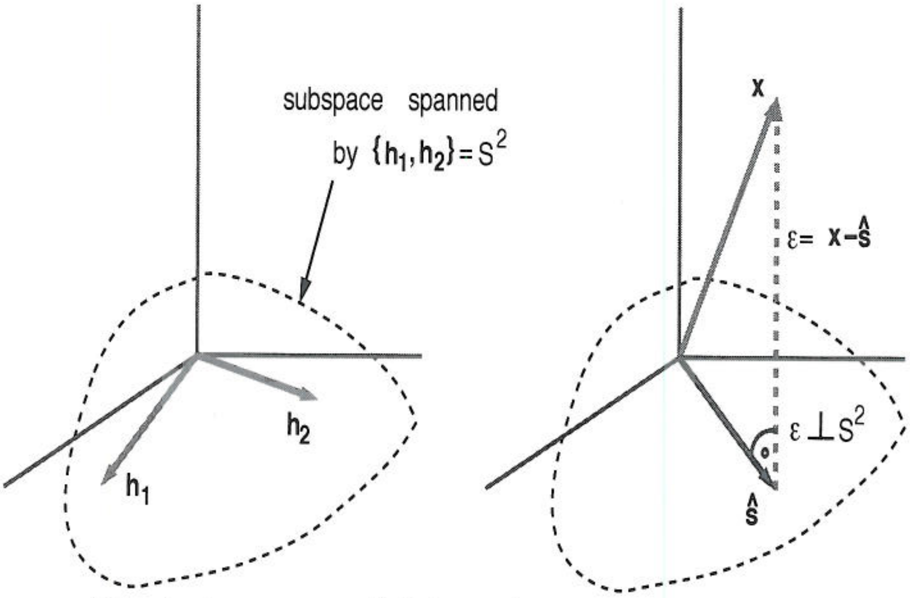
Set this result to zero to yield

$$\begin{aligned}(\mathbf{H}^T\mathbf{H})\boldsymbol{\theta} &= \mathbf{H}^T\mathbf{x} \\ \Rightarrow \qquad & \hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}\mathbf{x} \qquad \text{normal equations}\end{aligned}$$

b) [**combination of bookwork and application of theory**]
Geometric interpretations

$$\text{given} \quad \mathbf{s} = \mathbf{H}\boldsymbol{\theta} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} = \sum_{i=1}^{p} \theta_i \mathbf{h}_i$$

The vector $\mathbf{x} \in \mathbb{R}^N$, however, all signal vectors must lie in a p-dimensional subspace of $S^p \subset \mathbb{R}^N$. For example, for N=3, and $p = 2$, we have



a) Signal subspace b) Orthogonal projection to determine signal estimate

Signal vector model is the linear combination of the "signal" vectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_p\}$. The vector in S^2 which is closest to $\mathbf{x}$ in the Euclidean sense is the component $\hat{\mathbf{s}} \in S^2$, that is the "orthogonal projection" of $\mathbf{x}$ onto S^2. Two vectors in $\mathbb{R}^N$ are orthogonal if their scalar product $\mathbf{x}^T\mathbf{y} = 0$. Therefore, to determine $\hat{\mathbf{s}}$, we use the so-called orthogonality condition

$$(\mathbf{x} - \mathbf{s}) \perp S^2$$

or

$$\begin{aligned} A: &\quad (\mathbf{x} - \mathbf{s}) \perp \mathbf{h}_1 \;\Rightarrow\; (\mathbf{x} - \mathbf{s})^T \mathbf{h}_1 = 0 \\ B: &\quad (\mathbf{x} - \mathbf{s}) \perp \mathbf{h}_2 \;\Rightarrow\; (\mathbf{x} - \mathbf{s})^T \mathbf{h}_2 = 0 \end{aligned}$$

Using

$$\mathbf{s} = \theta_1 \mathbf{h}_1 + \theta_2 \mathbf{h}_2$$

and from the conditions A and B, we have

$$\begin{aligned}(\mathbf{x}-\theta_1\mathbf{h_1}-\theta_2\mathbf{h_2})^{\mathbf{T}}\mathbf{h_1} &= 0\\ (\mathbf{x}-\theta_1\mathbf{h_1}-\theta_2\mathbf{h_2})^{\mathbf{T}}\mathbf{h_2} &= 0\end{aligned}$$

which can be combined as

$$(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^{\mathbf{T}}\mathbf{H}=\mathbf{0}^{\mathbf{T}}$$

to yield the Least Squares Estimator (LSE)

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

where $\mathbf{H}$ is the measurement matrix.
Noting that

$$\boldsymbol{\varepsilon}=\mathbf{x}-\mathbf{H}\boldsymbol{\theta}$$

is the error vector, the error vector must be orthogonal to the observation matrix, i.e. $\boldsymbol{\varepsilon}^T\mathbf{H}=\mathbf{0}^T$. The error represents that part of $\underline{\boldsymbol{x}}$ which is not described by the signal model.

c) [**new example**]

We want to find the predictor coefficients $a_p[k]$ that minimise the linear prediction error $E_p=\sum_{n=0}^{\infty}(e[n])^2$.
To find these coefficients, differentiate E_p with respect to $a_p[k]$ and set the derivatives equal to zero as follows

$$\frac{\partial E_p}{\partial a_p[k]}=-\sum_{n=0}^{\infty}2e[n]\frac{\partial\hat{x}[n+n_0]}{\partial a_p[k]}=0$$

From $\hat{x}[n+n_0]=\sum_{k=1}^{p}a_p[k]x[n-k] \quad \Rightarrow \quad \frac{\partial\hat{x}[n+n_0]}{\partial a_p[k]}=x[n-k]$
Divide by two, and substitute for $e[n]$ to have,

$$\sum_{n=0}^{\infty}\left\{x[n+n_0]-\sum_{l=1}^{p}a_p[l]x[n-l]\right\}x[n-k]=0 \quad ; \quad k=1,2,\ldots,p$$

Therefore we obtain the normal equations

$$\sum_{l=1}^{p}a_p[l]r_x[k,l]=r_z[k,-n_0] \quad \textit{where} \quad r_z[k,l]=\sum_{n=0}^{\infty}x[n-l]x[n-k]$$

d) [**bookwork and above example**]

No probability assumptions are made about the data; only a signal model is assumed. Usually easy to implement, either in a block based or sequential manner, amounts to the minimisation of a least squares criteria. In the (LS) approach we attempt to minimise the squared difference between the observed data and the assumed signal or noiseless data.

5) a) [**bookwork**]
The input-output relation of the filter is given by

$$y = \sum_{k=1}^{p} w_k x_k$$

Let d denote the *desired response* or *target output* for the filter. Then the *error signal* is

$$e = d - y$$

As *performance measure* or *cost function*, we introduce the *mean squared error* defined as

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

The cost function

$$J = \frac{1}{2} E\{d^2\} - E\left\{\sum_{k=1}^{p} w_k x_k d\right\} + \frac{1}{2} E\left\{\sum_{j=1}^{p}\sum_{k=1}^{p} w_j w_k x_j x_k\right\}$$

and after some manipulation

$$J = \frac{1}{2} E\{d^2\} - \sum_{k=1}^{p} w_k E\{x_k d\} + \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p} w_j w_k E\{x_j x_k\}$$

Notation: $r_d = E\{d^2\}$
$r_{dx}(k) = E\{dx_k\}, \qquad k = 1, 2, \ldots, p$
$r_x(j,k) = E\{x_j x_k\}, \qquad j, k = 1, 2, \ldots, p$
Slot back into J to yield

$$J = \frac{1}{2} r_d - \sum_{k=1}^{p} w_k r_{dx}(k) + \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p} w_j w_k r_x(j,k)$$

To determine the optimum weights, follow the least squares approach

$$\nabla_{w_k} J = \frac{\partial J}{\partial w_k}, \qquad k = 1, \ldots, p$$

Differentiate wrt to w_k

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

and set to zero

$$\nabla_{w_k} J = 0, \qquad k = 1, 2, \ldots, p$$

Let w_{0k} denote the optimum setting of weight w_k. Then the optimum weights are determined by the following set of simultaneous equations

$$\sum_{j=1}^{p} w_{0j} r_x(j,k) = r_{xd}(k), \qquad k = 1, 2, \ldots, p$$

or in a compact form

$$\mathbf{w}_{opt} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{dx}$$

Within the method of steepest descent, the gradient of the error surface of the filter wrt the weights takes on a *time varying* form

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

The idea is to replace the block estimate from the Wiener filter with a recursive estimate on a much shorter data length. This allows for a sequential solution of this block filtering problem, which facilitates the use of short filters. However this way we introduce an error in the estimation.

Steepest decent: The weights have a **time–varying** form, they are adjusted in an **iterative** fashion along the error surface.

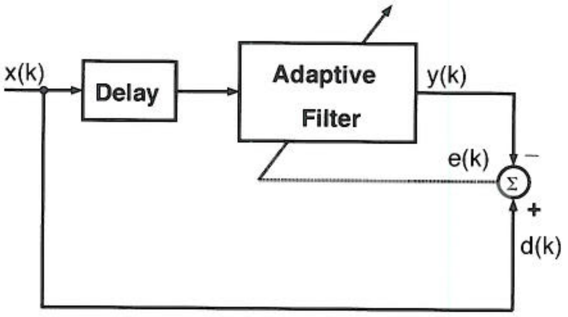
According to the method of steepest descent, the adjustment applied to the weight $w_k(n)$ at iteration n is defined by

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

where η is a positive constant called the **learning rate** parameter (step size).

b) [**bookwork**]
The adaptive prediction configurations is given in the figure below. The adaptive filter is fed with the delayed input x, and the teaching signal $d(n)$ is the current value of the input $x(n)$. The error $e(n)$ is used to adapt the filter coefficients.

11/11



Depending on the delay between the input and the teaching signal, this configuration can operate in a M step ahead prediction setting. The algorithms used to update the adaptive filter are the same as those used in any other adaptive filtering configuration.

c) [**new example**]
With $\mathbf{R}_x = \sigma_x^2\mathbf{I}$ and $\mathbf{r}_{dx} = \mathbf{0}$, the Wiener solution is $\mathbf{w} = \mathbf{0}$.

The steepest descent algorithm is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu[\mathbf{R}_x\mathbf{w}(n) - \mathbf{r}_{dx}]$$

Since $\mathbf{R}_x = \sigma_x^2\mathbf{I}$ and $\mathbf{r}_{dx} = \mathbf{0}$, then

$$\mathbf{w}(n+1) = \big(\mathbf{I} - \mu\sigma_x^2\mathbf{I}\big)\mathbf{w}(n)$$

With $\mu = 1/(5\sigma_x^2)$, the time evolution of $\mathbf{w}$ becomes

$$\mathbf{w}(n) = (1 - 1/5)^n\mathbf{w}_0$$

which asymptotically converges to $\mathbf{w}(\infty) = 0$, as $n \to \infty$.

d) [**new example**]
The update of the LMS algorithm is given by

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

Compared with the update of this variant of the Newton algorithm

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu e(n)\mathbf{R}_x^{-1}\mathbf{x}(n)$$

the stepsize μ within the LMS is replaced with $\mathbf{R}_x^{-1}\mu$, that is the input is decorrelated by the inverse of the correlation matrix and the algorithm converges to the optimal Wiener solution very fast. The contours of the error surface approach concentric circles, and the algorithm is not prone to gradient noise and zig-zagging to the optimum solution. In real world application, we can only estimate the true autocorrelation matrix, which will introduce gradient noise.