E4.13
AS2
SO15
ISE4.31

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2009

MSc and EEE/ISE PART IV: MEng and ACGI

Corrected Copy

SPECTRAL ESTIMATION AND ADAPTIVE SIGNAL PROCESSING

Wednesday, 20 May 10:00 am

Time allowed: 3:00 hours

There are FIVE questions on this paper.

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

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) : M.K. Gurcan, M.K. Gurcan

1) Consider the problem of periodogram based spectral estimation.

a) Explain the trade-off between the finite data length, maximum frequency in the signal, and the prescribed resolution and bandwidth. [2]

b) Bartlett's method partitions the input data $x[n]$, $n = 0, \ldots, N-1$ into K nonoverlapping sequences of length L, that is $N = K \times L$.

i) Given that the resolution of the standard periodogram is $\Delta\omega = 0.89\frac{2\pi}{N}$, what is the resolution of Bartlett's periodogram estimate? [2]

ii) Bartlett's method is used to estimate the power spectrum of a process from a sequence of $N = 2000$ samples. What is the minimum length L that may be used for each sequence if we are to have a resolution of $\Delta\omega = 0.005$. Explain why it would not be advantageous to increase L beyond the value found in the part $i)$ above. [4]

c) Many commercial Fourier analysers continuously update the estimate of the power spectrum of a process $x[n]$ by exponential averaging of periodograms, as follows

$$\hat{P}_i(f) = \lambda \hat{P}_{i-1}(f) + \frac{1-\lambda}{N}\left|\sum_{n=0}^{N-1} x_i[n]e^{-j\,2\pi fn}\right|^2, \quad f \in \left(-\frac{1}{2}, \frac{1}{2}\right]$$

where $x_i[n] = x[n + iN]$ is the i–th block of N data samples. The above update equation is initialised with $\hat{P}_{-1}(f) = 0, \quad \forall f$.

i) Assume that the periodogram is a random variable. Provide a detailed explanation of the fact that for stationary processes the above equation effectively performs low-pass filtering of the periodogram estimates, where λ is the coefficient of such a filter, and define the range for the parameter λ so that such an estimator is stable. [6]

ii) Assume the following notation

$$Q_i(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x_i[n]e^{-j2\pi fn}\right|^2$$

and express $\hat{P}_i(f)$ in terms of only $Q_k(f)$, $k = 0, \ldots, i$. Hence or otherwise, explain the effect of the windowing of the input data on the bias and variance of the estimator $\hat{P}_i(f)$ given above. [6]

2) Consider the problem of spectrum estimation of sinusoidal signals.

a) Spectral line splitting occurs when a single sinusoid in the spectrum is estimated as two sinusoids with frequencies close to the true frequency. Explain why parametric spectrum estimation methods are prone to this problem and suggest some ways to mitigate this effect. [4]

b) Explain the autocorrelation method for autoregressive spectrum estimation. [2]

c) The maximum entropy (ME) method is also based on autoregressive spectrum estimation. Provide a comparison of the autocorrelation method and the ME method. [4]

d) We would like to estimate the spectrum of a sinusoid using methods specifically designed for frequency estimation, which are based on the eigendecomposition of the autocorrelation matrix. For the estimation of one single frequency, the signal $x(n)$ and its autocorrelation $r_x(n)$ are given by

$$\begin{aligned} x(n) &= A_1 e^{jn\omega_1} + w(n) \\ r_x(n) &= P_1 e^{jk\omega_1} + \sigma_w^2 \delta(k) \end{aligned}$$

where $P_1 = |A_1|^2$ is the power of the complex exponential and $w(n)$ is white noise with variance σ_w^2.

i) If the useful signal and noise are uncorrelated, write down the expression for the autocorrelation matrix $\mathbf{R}_x$ in terms of the signal autocorrelation matrix $\mathbf{R}_s$ and the noise autocorrelation matrix $\mathbf{R}_n$. What is the rank of the signal autocorrelation matrix and the eigenstructure of the noise autocorrelation matrix? [4]

ii) Let $\mathbf{v}_i$ denote a noise eigenvector and $V_i(e^{j\omega}) = \sum_{k=0}^{M-1} v_i(k) e^{-j\omega k} = \mathbf{e}^H \mathbf{v}_i$. Write down the expression for a single frequency estimator based on this noise eigenvector. State the steps in the derivation of this method and explain whether such a spectrum is accurate over the whole frequency range. [6]

3) The coefficient update of the signed-regressor least mean square (LMS) algorithm is given by

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

where $e(k) = d(k) - y(k)$, $d(k), \mathbf{w}(k)$, and $\mathbf{x}(k)$ are respectively the desired response, coefficient vector and input vector, and the sign operator is applied to the vector $\mathbf{x}(k)$ component-wise.

a) Given Price's theorem which states that for a pair α and β of zero mean jointly Gaussian random variables, the statistical expectation $E\{\alpha\, sign(\beta)\}$ is given by

$$E\{\alpha\, sign(\beta)\} = \frac{1}{\sigma_\beta}\sqrt{\frac{2}{\pi}}\, E\{\alpha\beta\}$$

and any other necessary assumptions, which should be stated, show that

i)

$$E\{sign(\mathbf{x}(k))\mathbf{x}^T(k)\} = \frac{1}{\sigma_x}\sqrt{\frac{2}{\pi}}\,\mathbf{R} \quad \text{where} \quad \mathbf{R} = \mathrm{E}\{\mathbf{x}(\mathrm{k})\mathbf{x}^{\mathrm{T}}(\mathrm{k})\}$$

[4]

ii) The misalignment for the signed-regressor LMS algorithm $\mathbf{v}(k) = \mathbf{w}(k) - \mathbf{w}_{opt}(k)$ can be expressed as

$$E\{\mathbf{v}(k+1)\} = \left(\mathbf{I} - \mu\frac{1}{\sigma_x}\sqrt{\frac{2}{\pi}}\,\mathbf{R}\right) E\{\mathbf{v}(k)\}$$

where $\mathbf{w}_{opt}$ is the optimal Wiener weight vector. [6]

iii) Explain how the upper bound on the learning rate which preserves stability of this algorithm would be different from that for the LMS.

[4]

b) Consider a cost function given by

$$J(k) = |e(k)|$$

Show that the signed error LMS algorithm is a stochastic gradient solution based on the optimisation of this cost function. Write down the weight update equation for this algorithm and explain how you would derive a normalised signed error LMS algorithm.

(Hint: the derivative of $|e|$ is $sign(e)$). [6]

4) Consider the problem of finite impulse response (FIR) adaptive filtering.

a) A so called mixed norm cost function is given by

$$J(k) = \lambda|e(k)| + (1-\lambda)e^2(k)$$

where $e(k)$ is the instantaneous output error of the filter and $0 < \lambda < 1$ is a convex mixing parameter.

i) Derive a stochastic gradient based least mean square (LMS) type algorithm based on this cost function and compare the minimisation problem of LMS with this minimisation problem.
(Hint: the derivative of $|e(k)|$ is $sign(e(k))$. [4]

ii) Compare the behaviour of the mixed norm cost function for a small output error against its behaviour for a large output error. Hence provide the motivation for mixed norm adaptive filtering. [4]

iii) If the convex mixing parameter λ is made gradient adaptive, explain whether an adaptive filtering algorithm based on the mixed norm cost function can be realised via a hybrid filtering configuration. [4]

b) The outputs $\hat{x}(k+1)$ and $\hat{y}(k+1)$ of a dual channel (bivariate) real valued adaptive filter are given by

$$\begin{aligned} \hat{x}(k+1) &= \mathbf{a}^T(k)\mathbf{x}(k) + \mathbf{b}^T(k)\mathbf{y}(k) \\ \hat{y}(k+1) &= \mathbf{c}^T(k)\mathbf{x}(k) + \mathbf{d}(k)\mathbf{y}(k) \end{aligned}$$

where $\mathbf{a}(k), \mathbf{b}(k), \mathbf{c}(k)$ and $\mathbf{d}(k)$ are filter coefficient vectors, and $\mathbf{x}(k)$ and $\mathbf{y}(k)$ are the tap input vectors for the two input channels.

i) Based on the output errors $e_x(k) = x(k) - \hat{x}(k)$ and $e_y(k) = y(k) - \hat{y}(k)$, where $x(k)$ and $y(k)$ are teaching signals for the x and y channels, and the cost function

$$J(k) = \frac{1}{2}\left[e_x^2(k) + e_y^2(k)\right]$$

show that the LMS weight updates for the coefficient vectors $\mathbf{a}(k)$ and $\mathbf{b}(k)$ are given by [4]

$$\begin{aligned} \mathbf{a}(k+1) &= \mathbf{a}(k) + \mu e_x(k)\mathbf{x}(k) \\ \mathbf{b}(k+1) &= \mathbf{b}(k) + \mu e_x(k)\mathbf{y}(k) \end{aligned}$$

ii) Can this adaptive filter be realised using a collaborative (hybrid) filtering configuration? [4]

5) Consider an adaptive finite impulse response (FIR) filter.

a) Draw the block diagram of an adaptive prediction configuration and explain the operation of this adaptive filtering architecture. [3]

b) The adaptive prediction configuration can be used for adaptive line enhancement. Explain the relationship between the delay element within the adaptive line enhancement configuration and the correlation structure of the external noise source. [3]

c) A stochastic gradient weight update is given by

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \mu \nabla_{\mathbf{w}} J(k)$$

where $\mathbf{w}(k)$ are filter weights at time instant k and μ is the learning rate, a small positive constant. Derive the standard least mean square (LMS) algorithm based on the minimisation of the squared instantaneous output error $J = \frac{1}{2}e^2(k)$. Explain the role of the learning rate. [4]

d) The Least Mean Fourth (LMF) adaptive filtering algorithm is based on the minimisation of the cost function given by

$$J(k) = e^4(k)$$

i) Derive the LMS type weight update for this algorithm, based on the stochastic gradient update given in part c) of this question. In your own words comment on the sensitivity of this algorithm to the choice of the learning rate. [5]
(Hint: compare the steepness of the fourth order error surface of LMF and the second order error surface of LMS)

ii) The teaching signal for the LMF algorithm is given by

$$d(k) = \mathbf{x}^T(k)\mathbf{w}_{opt}(k) + q(k)$$

where $\mathbf{x}(k)$ is the input vector to the filter, $\mathbf{w}_{opt}$ is the optimal weight vector, and $q(k)$ is white Gaussian noise. Write down the minimum achievable mean square error J_{min} for the LMF algorithm and explain why it is the same as J_{min} for the LMS algorithm. [2]

ii) Describe in your own words the difference between the LMF and LMS algorithms. Which algorithm do you expect to perform better for a small error $e(k)$, and which for a large error? [3]

Master April 05 -

E4.13
S015
Ste 4.31
AS2

1/8

Solutions: Spectral Estimation 2009

1) a) [**bookwork**]

- Suppose we know the maximum frequency in the signal ω_{max}, and the required resolution $\Delta\omega$. Then

$$\Delta\omega > 2\frac{2\pi}{NT} \quad \Rightarrow \quad N > \frac{4\omega_{max}}{\Delta\omega}$$

- If we want to achieve both the prescribed resolution and bandwidth, then

$$\Omega_0 = \frac{1}{T} > 2\omega_{max} \quad \& \quad 2\omega_0 < \Delta\omega$$

$$\textit{hence}$$

$$\frac{\Omega_0}{2} = \frac{\pi}{T} > \omega_{max} \quad \textit{that is} \quad T < \frac{\pi}{\omega_{max}} \Leftrightarrow N > \frac{4\omega_{max}}{\Delta\omega}$$

- If we know the signal duration ($\frac{\Omega_0}{2} = \omega_{max} \rightarrow T < \frac{\pi}{\omega_{max}}$)

$$N > \frac{2t_{max}}{T} \quad \Rightarrow \quad N > \frac{2t_{max}\omega_{max}}{\pi}$$

$t_{max} \times \omega_{max} \rightarrow$ time–bandwidth product of a signal.

b) [**bookwork and new example**]
Bartlett's method aims at reducing the variance of the periodogram by the factor of K, at the expense of decreasing the resolution by the same factor. Thus the resolution $\Delta\omega = 0.89K\frac{2\pi}{N}$.

i) From a), $N = k \times L$, and the minimum length to achieve the resolution of $\Delta\omega = 0.005$ is calculated from the standard periodogram resolution for L data points, giving

$$L = 0.89\frac{2\pi}{\Delta\omega} \approx 2\pi * 180 \qquad \text{rounded}$$

Increasing L will increase the resolution but will also result in a decrease in the number of segments that may be averaged. This, in turn, will increase the variance of the spectrum estimate.

ii) c) [**new example**]
i) The goal with the spectrum analyser is to continuously refine the spectrum estimate as new data is read. With the arrival of each new data block, the periodogram is calculated and averaged with the previous spectrum estimate. Notice that due to the recursive nature of $\hat{P}_i$, it is suitable for mildly non-stationary processes. The choice $0 < \lambda < 1$ helps to forget the past value of $\hat{P}_i(f)$ as the new data is measured. When $\lambda = 0$ only the periodogram of the most recent data values is used, that is we have the standard periodogram estimator.

As the periodogram can be considered a random variable, say y, then transfer function of this periodogram estimator is

$$H(z) = \frac{\frac{1-\lambda}{N}}{1-\lambda z^{-1}}$$

which is consistent with the transfer function of a digital filter (lowpass or highpass). Clearly, $0 < \lambda < 1$ preserves stability. Therefore, we have a digital filtering operation, where the choice of λ determines whether the filter is stable or not.

ii) The power spectrum sequence is a geometric series of the random inputs $Q_i(f)$, hence

$$\hat{P}_i(f) = \sum_{k=0}^{i}(1-\lambda)\lambda^k Q_k(f)$$

and

$$E\{\hat{P}_i(f)\} = \sum_{k=0}^{i}(1-\lambda)\lambda^k E\{Q_k(f)\}$$

The windowing will not affect the bias of this periodogram estimator as all the windowed estimates are asymptotically unbiased. The resolution will however be window-dependent and the windowing will however help with the problem of spectral masking associated with the periodogram. The variance will however be

$$var\{\hat{P}_i(f)\} \approx (1-\lambda^{i+1})\frac{1}{2\pi}P_i(f)$$

2) [**bookwork and worked example**] a) Figure 2.1 illustrates the problem of spectral line splitting which is associated with the autoregressive spectrum estimation based on the autocorrelation method.

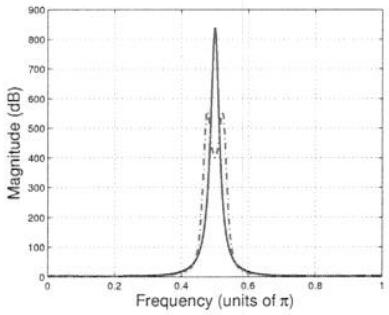

Figure 2.1. Solid line - spectrum estimate of a sinewave. Broken line - spectral line splitting

Spectral line is a phenomenon which is typical for overmodelled parametric spectrum estimates (the estimated model order p is too large). It therefore introduces false peaks in the spectra which can mistake e.g. a single sinusoid for two sinusoid with close frequencies. This effect can be mitigated by more accurate estimates of the AR coefficients and the order of the underlying AR model. This can be achieved, for instance, by the autocovariance, modified autocovariance, or Burg's method.

b) An AR(p) process may be represented as the output of an all–pole filter driven with unit variance white noise. Its spectrum is given by

$$P_{xx}(\omega) = \frac{|b(0)|^2}{\left|1 + \sum_{k=1}^{p} a_p(k) e^{-\jmath k\omega}\right|^2}$$

Since $b(0)$ and $a_p(k)$ can be estimated from the data, an estimate of the power spectrum becomes

$$\hat{P}_{AR}(\omega) = \frac{|\hat{b}(0)|^2}{\left|1 + \sum_{k=1}^{p} \hat{a}_p(k) e^{-\jmath k\omega}\right|^2}$$

The AR coefficients are found by solving the ACF normal equations

$$\begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(p) \\ r_x(1) & r_x(0) & \cdots & r_x(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(p) & r_x(p-1) & \cdots & r_x(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_p(1) \\ \vdots \\ a_p(p) \end{bmatrix} = \varepsilon_p \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where

$$r_x(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x(n+k)x(n) \qquad k = 0, \ldots, p$$

and ε_p is the driving noise variance.

c) In the Yule–Walker method $x(n)$ is assumed to be AR process, whereas in the maximum entropy method it is assumed that $x(n)$ is Gaussian. The only real difference is in the assumptions imposed on process $x(n)$: the Yule–Walker method assumes $x(n)$ is an autoregressive process, whereas the ME method assumes $x(n)$ is Gaussian as it extrapolates the autocorrelation function based on maximum entropy. The ACF is effectively extrapolated with zeros, hence the ACF method generally produces low resolution estimates than the approaches that do not window the data. Consequently, this method is generally not used for short data records. d) i) Since $s \perp n$ (subspaces analysis – later), $\mathbf{R}_{xx}$ becomes rank one

$$\mathbf{R}_{xx} = \mathbf{R}_s + \mathbf{R}_n \qquad \mathbf{R}_s = \begin{bmatrix} 1 & e^{-\jmath\omega_1} & \dots & e^{-\jmath(M-1)\omega_1} \\ e^{\jmath\omega_1} & 1 & \dots & e^{-\jmath(M-2)\omega_1} \\ \vdots & \vdots & \ddots & \vdots \\ e^{\jmath(M-1)\omega_1} & e^{-\jmath(M-2)\omega_1} & \dots & 1 \end{bmatrix}$$

and

$$\mathbf{R}_n = \sigma_w^2 \mathbf{I}$$

ii) Define $\mathbf{e}_1 = \left[1, e^{\jmath\omega_1}, \dots, e^{\jmath(M-1)\omega_1}\right]^T$, thus $\mathbf{R}_s = |A_1|^2 \mathbf{e}_1 \mathbf{e}_1^H$ is rank one and has one nonzero eigenvalue $M|A_1|^2$.
$\mathbf{R}_s$ is Hermitian and the remaining eigenvectors $\mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_M$ are orthogonal to $\mathbf{e}_1$, that is $\mathbf{e}_1^H \mathbf{v}_i = 0 \qquad i = 2, 3, \dots, M$.
Notice also that $\mathbf{R}_{xx}\mathbf{v}_i = \left(\mathbf{R}_s + \sigma_w^2 \mathbf{I}\right) \mathbf{v}_i = \left(\lambda_i^s \mathbf{v}_i + \sigma_w^2 \mathbf{v}_i\right) = \left(\lambda_i^s + \sigma_w^2\right) \mathbf{v}_i$

- $\Rightarrow$ the eigenvectors of $\mathbf{R}_{xx}$ are the same as those of $\mathbf{R}_s$ and the eigenvalues of $\mathbf{R}_{xx}$ are $\lambda_i = \lambda_i^s + \sigma_w^2$
- The largest eigenvalue of $\mathbf{R}_{xx}$ is $\lambda_{max} = M|A_1|^2 + \sigma_w^2$, and the remaining $(M-1)$ eigenvalues are equal to σ_w^2

We then have to perform the eigenvalue decomposition and take the largest eigenvalue as the signal eigenvalue and the rest as noise eigenvalues. We can then determine the frequency ω_1 from the eigenvector $\mathbf{v}_{max}$ that is associated with the largest eigenvalue using. Thus frequency estimation can be performed using

$$\hat{P}_i(\omega) = \frac{1}{\left|\sum_{k=0}^{M-1} v_i(k) e^{-\jmath k\omega}\right|^2} = \frac{1}{|\mathbf{e}^H \mathbf{v}_i|^2}$$

Such a spectrum estimate is accurate only in determining the frequency of the sinusoids, however, it is not accurate over the whole frequency range.

3) a) [**new example**]

i) Assume that the elements of $\mathbf{x}(k)$ are zero mean Gaussian random variables. As

$$sign(\mathbf{x}(k)) = [sign(x(k)), sign(x(k-1)), \ldots, sign(x(k-N+1))]^T$$

the product

$$sign(\mathbf{x}(k))\mathbf{x}^T(k)) = \begin{bmatrix} sign(x(k))x(k) & \cdots & sign(x(k))x(k-N+1) \\ \vdots & \ddots & \vdots \\ sign(x(k-N+1))x(k) & \cdots & sign(x(k-N+1))x(k-N+1) \end{bmatrix}$$

From Price's theorem, $E\{sign(\mathbf{x}(k))\mathbf{x}^T(k)\} = \frac{1}{\sigma_x}\sqrt{\frac{2}{\pi}}\, E\{\mathbf{x}(k)\mathbf{x}^T(k)\} = \frac{1}{\sigma_x}\sqrt{\frac{2}{\pi}}\, \mathbf{R}$.

ii) Similarly to the convergence in the mean of the LMS algorithm, we have

$$\begin{aligned} \mathbf{w}(k+1) &= \mathbf{w}(k) + \mu sign(\mathbf{x}(k))(d(k) - \mathbf{x}^T(k)\mathbf{w}(k)) \\ &= \mathbf{w}(k) - \mu sign(\mathbf{x}(k))\mathbf{x}^T(k)\mathbf{w}(k) + \mu d(k) sign(\mathbf{x}(k)) \\ &= \left(\mathbf{I} - \mu sign(\mathbf{x}(k))\mathbf{x}^T(k)\right)\mathbf{w}(k) + \mu d(k) sign(\mathbf{x}(k)) \end{aligned}$$

From the Wiener solution, subtract $\mathbf{w}_{opt}$ from the above equation to obtain

$$E\{\mathbf{w}(k+1) - \mathbf{w}_{opt}\} = \left(\mathbf{I} - \mu\frac{1}{\sigma_x}\sqrt{\frac{2}{\pi}}\,\mathbf{R}\right) E\{\mathbf{w}(k)\} - \left(\mathbf{I} - \mu\frac{1}{\sigma_x}\sqrt{\frac{2}{\pi}}\,\mathbf{R}\right)\mathbf{w}_{opt}$$

Since $\mathbf{v}(k+1) = \mathbf{w}(k) - \mathbf{w}_{opt}$, from the above equation we directly have the desired expression.

iii) The bounds on the learning rate are obtained from the modes of convergence, that is, based on the reciprocal of the largest eigenvalue of the correlation matrix. Based on Price's theorem, it is straightforward to show that the maximum learning rate will be corrected by the factor $\sigma_x\sqrt{\frac{2}{\pi}}$.

b) Similarly to the LMS, to calculate the update

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \mu\nabla_{\mathbf{w}}J(k) = \mathbf{w}(k) + \mu sign(e(k))\mathbf{x}(k)$$

we have

$$\nabla_{\mathbf{w}}J(k) = \frac{|e(k)|}{\partial e(k)}\frac{\partial e(k)}{\partial y(k)}\frac{\partial y(k)}{\partial \mathbf{w}(k)} = -sign(e(k))\mathbf{x}(k)$$

This filter can be normalised by minimising the a posteriori prediction error $e(k+1)$ using the Taylor series expansion around $e(k)$.

4) a) [**bookwork and worked example**]
i) Similarly to the answer 3c), we can combine the sign error and LMS type update to obtain

$$\underline{\mathrm{w}}(k+1) = \mathbf{w}(k) + \lambda sign(e(k))bfx(k) + 2(1-\lambda)e(k)\mathbf{w}(k)$$

This is basically an LMS type update which should provide a compromise between the algorithms based on the minimisation of the first and second order cost function. (You should provide a detailed derivation in the answer).
ii) For a small error, the term $e^2(k)$ become even smaller and the weight update is dominated by the sign error algorithm. For a large error, e^2 is the dominant term. If we know the statistics of the input signal, then we can choose λ for an optimal performance which is better than the performance of the individual filters based on $e(k)$ and $e^2(k)$.

iii) The adaptive $\lambda(k)$ would help to cope with the time varying statistics of the input. Hybrid filters are based on a similar convex combination, based on $\lambda(k)$. However, within a hybrid filter we combine two different adaptive filters with two different weight vectors, which are updated separately based on their own cost function. The mixed norm solution uses a combination of cost functions, however, the mixed norm filter has only one weight vector and cannot be realised as a hybrid filter.

b) **new example**
Based on standard LMS algorithm, we have

$$\begin{bmatrix} \mathbf{a}(k+1) \\ \mathbf{b}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{a}(k) \\ \mathbf{b}(k) \end{bmatrix} + \mu e_x(k) \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{y}(k) \end{bmatrix}$$
$$\begin{bmatrix} \mathbf{c}(k+1) \\ \mathbf{d}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{d}(k) \\ \mathbf{d}(k) \end{bmatrix} + \mu e_y(k) \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{y}(k) \end{bmatrix}$$

This filter can be realised as a hybrid filter, as the inputs to both x and y channel are both $\mathbf{x}(k)$ and $\mathbf{y}(k$, and the channel coefficients are updated based on their own instantaneous error. For a proper hybrid filter, the cost function would have to comprise the mixing parameter $\lambda(k)$.

5) a) [**bookwork**]
This is a standard adaptive filtering configuration where the teaching signal is

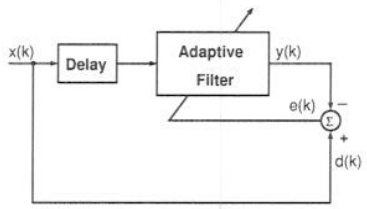

Figure 1: Left: Adaptive prediction configuration

the input signal advanced in time, so it can be used to train the filter to predict future values of the input. This is a core adaptive filtering configuration and can be used to predict the input M steps ahead, $M-1, 2, \ldots$.

b) [**bookwork and intuitive reasoning**]

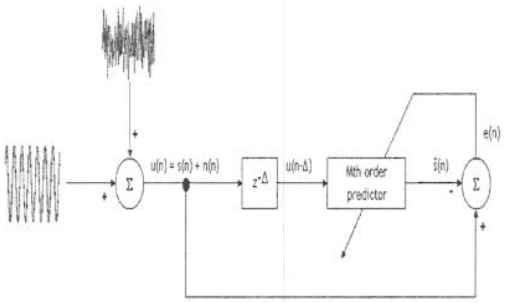

Figure 2: Adaptive Line Enhancement configuration

This architecture is similar to the adaptive prediction architecture, it the sense that the input and the teaching signal are the same signal delayed (advanced) in time (but for the noise in the input to the adaptive filter). The amount of delay depends on the correlation structure of the noise source. For a white additive noise, ideally we would need the delay by only one time sample. For correlated noises this delay should be long enough for the correlation function of the noise to decay significantly.

c) [**bookwork**]
Starting from the cost function $J(k) = \frac{1}{2}e^2(k)$, the update of the LMS algorithm is $\Delta\mathbf{w}(k) = \mu e(k)\mathbf{x}(k)$. It can be obtained from the cost function $J(k)$

and based on $e(k) = d(k) - y(k)$, using the chain rule, as

$$\nabla_{\mathbf{w}} J(k) = \frac{\partial \frac{1}{2} e^2(k)}{\partial e(k)} \frac{\partial e(k)}{\partial y(k)} \frac{\partial y(k)}{\partial \mathbf{w}(k)} = -e(k)\mathbf{x}(k)$$

to give

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

The learning rate controls the speed of adaptation and its optimal value depends on the correlation structure of the input. Ideally, it should be large in the beginning of adaptation (for fast convergence), and small towards the end of adaptation (for small misadjustment).

d) [**new example**] i) The LMF update can be derived in the same way as the LMS in c), that is based on the gradient of the cost function as follows

$$\nabla_{\mathbf{w}} J(k) = \underbrace{\frac{\partial e^4(k)}{\partial e(k)}}_{4e^3(k)} \underbrace{\frac{\partial e(k)}{\partial y(k)}}_{-1} \underbrace{\frac{\partial y(k)}{\partial \mathbf{w}(k)}}_{\mathbf{x}(k)} = -4e^3(k)\mathbf{x}(k)$$

to give

$$\mathbf{w}(k+1) = \mathbf{w}(k) + 4\mu e^3(k)\mathbf{x}(k)$$

The relationship between the learning rate and the correlation structure of the signal is more complicated than for LMS, as to arrive at the modes of convergence we would have the third power of the error and the associated terms. This algorithm is therefore likely to be more sensitive to the changes in the learning rate than the LMS. This can be shown graphically by comparing a fourth order error surface given by $e^4(k)$ and the second order error surface given by $e^2(k)$. The fourth order surface is much steeper, hence the LMF is more sensitive to the choice of the learning rate.

ii) The minimum achievable mean square error $J_{min} = \sigma_q^2$, as this is how close we can approach the teaching signal. This does not depend on the cost function, and depends purely on the Wiener solution. This quantity is therefore the same as in the case of LMS.

iii) For small errors $e^3(k) << e(k)$ and for large errors $e^3(k) >> e(k)$. This algorithm should therefore provide fine estimates when close to the optimum weights (in the end of adaptation), but can however be rather unstable for large errors (beginning of adaptation).