EE4-13
EE9SO15

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2011

MSc and EEE/ISE PART IV: MEng and ACGI

SPECTRAL ESTIMATION AND ADAPTIVE SIGNAL PROCESSING

Thursday, 12 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 nonparametric spectrum estimation.

a) Write down the expression for the periodogram and derive this expression starting from the Discrete Fourier Transform (DFT) of a discrete time signal $x(n)$. [4]

b) Explain the properties of the periodogram as an estimator of power spectral density (bias, variance, bias-variance tradeoff). [3]

c) Comment on the performance of the periodogram for peaky spectra, smooth spectra, and for two closely spaced sinewaves in noise. [3]

d) Consider a single sinusoid with angular frequency ω_0 contaminated with white noise $w(n) \sim \mathcal{N}(0, \sigma_w^2)$, given by

$$x(n) = Asin(n\omega_0 + \phi) + w(n)$$

i) The first three values of the autocorrelation sequence are known, and are given by

$$r_x(0) = 1 \quad r_x(1) = \beta \quad r_x(2) = 0$$

Find and prepare a carefully labeled sketch of the spectrum estimate that is formed using the Blackman–Tukey method with a rectangular window. [4]

ii) We desire to compute the periodogram $\hat{P}_{per}(e^{j\omega})$ using N samples of $x(n)$ (Bartlett method). Prepare a carefully labeled sketch of the expected value of this spectrum estimate. Also sketch the periodogram estimate of two closely spaced sinewaves contaminated with white Gaussian noise. Comment on the properties of this estimator in terms of its bias and variance. [6]

2) Modern spectrum estimation methods (Pisarenko, MUSIC) usually assume a harmonic model of the useful signal, that is, a single sinewave or a sum of sinewaves contaminated by noise.

a) Explain the principles that enable this class of techniques to extract the correct spectrum of a single sinusoid in noise. Comment on the accuracy of the spectrum estimate outside the frequency of the useful signal. [6]

b) The 'Principal Component' spectrum estimation method is based on the eigendecomposition of the autocorrelation matrix $\mathbf{R}_{xx}$, given by

$$\mathbf{R}_{xx} = \sum_{i=1}^{M} \lambda_i \mathbf{v}_i \mathbf{v}_i^H = \underbrace{\sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^H}_{signal\ subspace} + \underbrace{\sum_{i=p+1}^{M} \lambda_i \mathbf{v}_i \mathbf{v}_i^H}_{noise\ subspace}$$

where $\lambda_i, i = 1, \ldots, M$ are the eigenvalues and $\mathbf{v}_i$ the eigenvectors of $\mathbf{R}_{xx}$.

i) Explain whether this method is more robust than methods based on the noise subspace (e.g. Pisarenko), and write down the expression for power spectrum estimate based on the signal subspace within the above decomposition. Which method would you prefer for the estimation of multiple sinewaves in white Gaussian noise? [8]

ii) We may use this method in combination with other modern spectrum estimation methods. Explain the principles and benefits of combining it with autoregressive spectrum estimation. [2]

c) We desire to estimate the power spectrum of the autoregressive process of order two, $AR(2)$, given by

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

where $w[n] = \mathcal{N}(0, 1)$. The measurements $y[n]$ of the useful signal $x[n]$ are corrupted by noise, and we observe

$$y[n] = x[n] + v[n]$$

where $v[n]$ a moving average (MA) process that is uncorrelated with $x[n]$, that is $v[n] \perp x[n]$. The corresponding autocorrelation functions are

$$r_{yy}[0] = 5 \quad r_{yy}[1] = 2 \quad r_{yy}[2] = 0 \quad r_{yy}[3] = -1 \quad r_{yy}[4] = 0.5$$
$$r_{vv}[0] = 3 \quad r_{vv}[1] = 1$$

Explain in your own words (give some mathematical support) a way to estimate the power spectrum of the useful signal $x[n]$. [4]

(Hint: for uncorrelated processes $r_{yy}[n] = r_{xx}[n] + r_{vv}[n]$)

3) Consider the problem of multistep prediction for a random process having an autocorrelation function

$$r_{xx}(k) = \delta(k) + 0.9^{|k|}\cos(\frac{\pi k}{4})$$

for which the first eight values are

$$\mathbf{r}_{xx} = [2.0, 0.6364, 0, -0.5155, -0.6561, -0.4175, 0, 0.3382]^T$$

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

b) Use the Wiener filter to predict the above signal.

i) Find the coefficients of the second-order Wiener filter acting as a one-step ahead predictor, and write down the equation for such a predictor. Comment on the teaching signal used in the prediction configuration, and write down the expression for the minimum mean-square error. [5]

Hint: The solutions is based on $\begin{bmatrix} r_{xx}(0) & r_{xx}(1) \\ r_{xx}(1) & r_{xx}(0) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} r_{xx}(1) \\ r_{xx}(2) \end{bmatrix}$

ii) Repeat part i) for a three-step ahead predictor. Comment on the minimum mean-square error obtained in this case. [5]

c) Consider an adaptive predictor for both the cases in part b). Write down the input-output equations of these two predictors and derive the Least Mean Square (LMS) type of update based on the minimisation of the instantaneous output error $J = \frac{1}{2}e^2(n)$. [6]

4) Consider the problem of complex-valued adaptive filtering.

a) Draw the block diagram of a widely-linear adaptive filter and explain in your own words the need for widely linear modelling. [6]

b) The two gradients obtained in the optimisation of complex-valued error surfaces are $\nabla_{\mathbf{w}} J$ and $\nabla_{\mathbf{w}^*} J$. Write down the expressions for the $\mathbb{R}$ and $\mathbb{R}^*$ derivatives and explain which of the two gradients provides the steepest descent along the error surface. [6]

c) The Constant Modulus Algorithm (CMA) operates on complex-valued processes with a constant envelope, e.g. phase modulated signals. The output of such a filter is

$$y(n) = \mathbf{w}^H(n)\mathbf{x}(n)$$

and the error

$$e(n) = \frac{1}{2}\left(|y(n)|^2 - 1\right) \qquad \Rightarrow \qquad J(n) = \frac{1}{4}E\left\{\left(|y(n)|^2 - 1\right)^2\right\}$$

is real and non-negative. Replacing the statistical averages in the cost function $J(n)$ by the instantaneous estimates, derive the CMA algorithm, given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\left(1 - |y(n)|^2\right)y^*(n)\mathbf{x}(n)$$

which is a Least Mean Square (LMS) type algorithm applied to $J(n)$. [8]

Hint: Use the fact that

$$|y(n)|^2 = \mathbf{w}^H(n)\mathbf{x}(n)\mathbf{x}^H(n)\mathbf{w}(n) \qquad \& \qquad \frac{\partial}{\partial \mathbf{w}^H}\mathbf{w}^H\mathbf{A}\mathbf{w} = \mathbf{A}\mathbf{w}$$

5) This questions addresses different criteria for the minimisation of the error performance surface (cost function).

a) A family of stochastic gradient algorithms is based upon approximately minimising the cost function of the form

$$J = E\left\{e^{2p}(n)\right\}, \quad p = 1, 2, \ldots$$

where $e(n) = d(n) - y(n)$, namely the difference between the desired response $d(n)$ and the output of the adaptive filter $y(n) = \mathbf{x}^T(n)\mathbf{w}(n)$, where $\mathbf{w}(n) = [w_1(n), \ldots, w_N(n)]^T$ is the coefficient (weight) vector of an N–tap, finitie impulse response, adaptive filter with input vector $\mathbf{x}(n) = [x(n), x[(n-1), \ldots, x(n-N+1)]^T$.
Verify that a least mean square (LMS) type coefficient update for $\mathbf{w}(n)$, based upon J, is given by

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\,p\,\mu\,e^{2p-1}(n)\mathbf{x}(n) \quad .$$

and comment on the sensitivity of this class algorithms for small and large values of p. [6]

b) Now consider the 'mixed norm' cost function

$$J = \alpha|e(n)| + (1-\alpha)\frac{1}{2}e^2(n), \qquad 0 \le \alpha \le 1$$

Derive an LMS-type algorithm based on this cost function, and comment on its potential applications. [8]

c) Consider the cost function

$$J(n) = \frac{1}{2}\exp\left(\mathrm{e}^2(\mathrm{n})\right) = \frac{1}{2}\mathrm{e}^{\mathrm{e}^2(\mathrm{n})}$$

Derive an LMS-type algorithm based on this cost function, write down a simplified form of the algorithm for small output errors, and explain in your own words the similarities with the algorithms in a) and b). [4]

d) Explain which of the above algorithms would perform best on inputs contaminated with impulsive additive noise (large output errors). [2]

Hint: $\frac{d|x(n)|}{dx} = \mathrm{sign(x)} = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$ $\qquad \mathrm{e^x} = 1 + \mathrm{x} + \frac{\mathrm{x}^2}{2!} + \frac{\mathrm{x}^3}{3!} + \cdots$

EE4-13
EE9 2015

Spectral Estimation & Adaptive Signal Processing 1/8 10

Solutions: 2011

1) a), b), and c) [**bookwork**]
Start from the Fourier transform of the autocorrelation function

$$P_x(e^{\jmath\omega}) = \sum_{k=-\infty}^{\infty} r_x(k)e^{-\jmath k\omega}$$

There are several ways to show its relation with the DFT, one intuitive way is

$$x_N \rightarrow X_N(k) \rightarrow \frac{1}{N}|X_N(k)|^2 = \hat{P}_{per}(e^{\jmath 2\pi k/N})$$

When considering the periodogram as an estimator of power spectral density, we need to address

$$E\{r_x(k)\} = \frac{N-|k|}{N} r_x(k) \Rightarrow E\{\hat{P}_{per}(e^{\jmath\omega}\} = \frac{2}{\pi} P_x(e^{\jmath\omega}) * W_B(e^{\jmath\omega})$$

where W_B is the Fourier transform of Bartlett window. Since the above estimate of the autocorrelation function is biased, the periodogram is a biased estimator, but since W_B converges to an impulse as N goes to infinity, it is asymptotically unbiased.

When it comes to the variance of the periodogram, we use the standard definition of varience of an estimator. Due to the complicated expression for the periodogram, it is mathematically most tractable to consider its variance for a white input. In that case, we can use the 'fourth order moment separation' theorem, to arrive at

$$var\{\hat{P}_{per}(e^{\jmath\omega})\} = P_x^2(e^{\jmath\omega})$$

Periodograms are MA-type estimators and are hence not suitable for peaky spectra. They are much better suited for smooth spectra (MA spectra). For two closely spaced sinewaves in noise, the periodogram is not the best choice, as the artifacts due to the finite observation window (main and sidelobes of the sinc function) can mask the spectral contents.

d) **new example**
i) You need to show (requires some calculation) that

$$\hat{P}_x(e^{\jmath\omega}) = \sum_{k=-M}^{M} r_x(k)e^{-\jmath\omega} = 1 + 2\beta cos\omega$$

ii)Start from

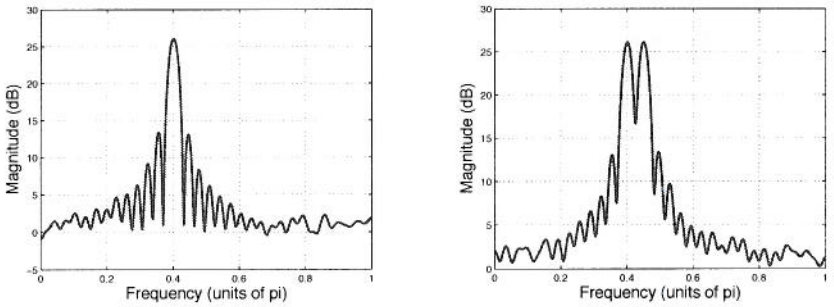
$$E\{\hat{P}_{per}(e^{\jmath\omega}\} = \frac{1}{2\pi} P_x(e^{\jmath\omega}) * W_B(e^{\jmath\omega})$$

$$W_B(e^{\jmath\omega}) = \frac{1}{N}\left[\frac{sin(N\omega/2)}{sin(\omega/2)}\right]^2$$

$$P_x(e^{\jmath\omega}) = 1/2\pi A^2[u_0(\omega-\omega_0) + u_0(\omega+\omega_0)] + \sigma_w^2$$

$$\Rightarrow E\{\hat{P}_{per}(e^{\jmath\omega}\} = \frac{2}{\pi} P_x(e^{\jmath\omega}) * W_B(e^{\jmath\omega}) =$$

$$= \sigma_w^2 + 1/4A^2[W_B(e^{\jmath(\omega-\omega_0)}) + W_B(e^{\jmath(\omega+\omega_0)}]$$



(a) Power spectrum of sinewave in noise (b) Power spectrum of two sinewaves in noise

Figure 1: Power spectra of noisy sinewaves

The periodogam is biased and since the variance does not go to zero as $N \to \infty$ it is not a consistent estimate ot the power spectrum.
You effectively have a noise floor and two 'sinc' functions at the locations of true sinewaves (due to the convolution in frequency of the spectrum of rectangular window and true spectrum). The graphs for one and two sinewaves in noise are shown below Care should be taken wrt the resolution (dictated by the number of datapoints N). For two closely spaced sinewaves, we need to inssure that the mainlobe of the *sinc* is narrow enough to be able to discriminate between the two sinewaves, and at the same time that the sidelobes are suppressed enough in order not to mask a possibly weak sinewave.

2) a) [**bookwork**]
These methods assume that power spectum at a discrete set of frequencies has physical meaning (information bearing such is in radar, sonar, speech, biomedical engineering). The idea is to use eigendecomposition to decompose the autocorrelation matrix of the (noisy) data into the signal-related part and noise-related part. Based on the orthogonality between the useful signal and noise, and the orthogonality between the eigenvectors in eigenvalue decomposition, we may use either the noise subspace or signal subspace for power spectrum estimation at the desired set of frequencies of interest. For instance, noise-subspace methods make use of the orthogonality of signal eivenvectors and noise eigenvectors, that is (for p sinusoids in noise and N observations)

$$\mathbf{e_i}^H \mathbf{v_j} = 0 \qquad i = 1, \ldots, p \quad j = p+1, \ldots, N$$

where

$$\mathbf{e}_i = [1, e^{-j\omega_i}, \ldots, e^{-jp\omega_i}]^T$$

is the vector of complex exponentials corresponding to the desired sinewave frequency ω_i, and $\mathbf{v}_j$ an eigenvector belonging to the noise subspace. We can then simply use the following expression as the spectrum estimator

$$\frac{1}{e_i^H \mathbf{v}_j}$$

as it will give a peak at the frequency of the desired sinewave ω_i. Combining all the p such estimators gives a more robust estimate. The so produced power spectrum estimate need not be accurate outside the discrete set of frequencies of interest.

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

$$\hat{R}_s = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^H$$

The methods based on the noise subspace estimation produce a peak in the spectrum for a discrete set of frequencies of interest. The PCA based spectrum estimate on the other hand imposes a rank constraint on the signal subspace and provides an estimate of the ACF of the signal.
Since this method produces an estimate of ACF, it can be used in conjunction with other standard methods which rely on an estimate of ACF.
The choice of the method depends on the dimensionality of the problem. For a few sinewaves and many datasamples, the noise subspace has a large dimension and the noise-subspace based estimates can be avereged to produce robust estimation. Alternatively, for many sinewaves and not many datapoints, it is more convenient to use PCA to reduce the signal space and then use e.g. autoregressive or maximum entropy spectrum estimation to obtain more accurate results.

iv) $\mathbf{R}_{ss}$ from above can be used directly within autoregressive spectrum estimation since it provides an estimate of ACF. It can also be used within MEM, due to the duality between the autoregressive and MEM spectrum estimation.

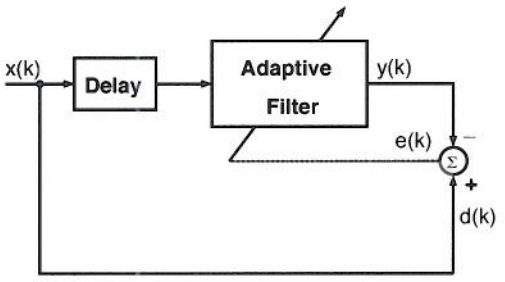
c) [**new example**]
Due to the orthogonality between signal and noise and the MA noise model we have

$$\begin{aligned} r_{xx}[n] &= r_{yy}[n] - r_{vv}[n] \quad \text{which gives} \\ r_{xx}[0] = 2 \qquad & r_{xx}[1] = 1 \quad r_{xx}[2] = 0 \quad r_{xx}[3] = -1 \quad r_{xx}[4] = 0.5 \end{aligned}$$

We therefore know the dimension of the signal subspace and by performing since we know the ACF of the data, we can employ any ACF based spectrum estimation method.

5

3) a) **[bookwork]**
The adaptive prediction configuration is shown in the Figure below. The input



signal x is also acting as a teaching signal, when advanced by the required number of steps in time. For practical purposes this is denoted by the delay in the direct branch to the filter. Since the teaching signal is advanced input, in the Wiener filter we only need to use the autocorrelations. The adaptive filtering block can use any of general adaptive filtering algorithms.

b) [**bookwork and new example**]
i) Starting from a second-order predictor in the form (note that for the Wiener filter the weights are fixed)

$$\hat{x}(n+1) = w_1 x(n) + w_2 x(n-1)$$

and solving for the filter coefficients using the solution given in the Hint, we have

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0.3540 \\ -0.1127 \end{bmatrix}$$

giving the optimum second-order one-step ahead predictor in the form

$$\hat{x}(n+1) = 0.3540x(n) - 0.1127x(n-1)$$

The minimum mean-square error is obtained from (having in mind that the teaching signal $d(n) = x(n+1)$)

$$J_{min} = E\{e^2(n)\} = E\{(\hat{x}(n) - x(n))^2\} = r_{xx}(0) - w_1 r_{xx}(1) - w_2 r_{xx}(2) = 1.7747$$

ii) For a third order predictor we have

$$\hat{x}(n+3) = w_1 x(n) + w_2 x(n-1)$$

and the Wiener-Hopf equations become

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(1) \\ r_{xx}(1) & r_{xx}(0) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} r_{xx}(3) \\ r_{xx}(4) \end{bmatrix}$$

leading to the set of coefficients

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -0.1706 \\ -0.2738 \end{bmatrix}$$

and the optimum predictor in the form

$$\hat{x}(n+3) = -0.1706x(n) - 0.2738x(n-1)$$

$$\mathbf{x}^T(k)\mathbf{w}(k+1) = \mathbf{x}^T(k)\mathbf{w}(k) + \alpha(k)\mathbf{x}^T(k)\mathbf{k}(k)$$

The minimum mean-square error becomes

$$J_{min} = r_{xx}(0) - w_1 r_{xx}(3) - w_2 r_{xx}(4) = 1.7324$$

Unexpectedly, it is smaller than that for the one-step ahead prediction. Tis can be explained with the high values of the correlation function for lags 3 and 4, with the correlation for lag 4 being larger than the correlatino for lag 1.

c) **bookwork and new example**
Starting from $J = 1/2e^2(n)$, we have

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

Insert into the steepest descent equation to obtain the LMS

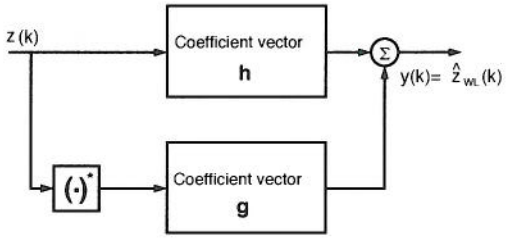
$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu\nabla_{\mathbf{w}}J = \mathbf{w}(n) + \mu e(n)\mathbf{x}(n)$$

In both the prediction cases the input-output equations are

$$y(n) = w_1(n)x(n) + w_2(n)x(n-1)$$

and the standard LMS, derived above can be used to update the filter weights. The choice between a one-step ahead and three-step ahead prediction is achieved by appropriately choosing the teaching signal, as explained in part a).

4) **Bookwork and new example**
The block diagram of the widely linear filter is given in the Figure below.



It has a direct branch with the filter coefficent vector $\mathbf{h}$ and the conjugate branch with the filter coefficent vector $\mathbf{g}$, and thus satisfies the widely linear model

$$y(n) = \mathbf{h}^T(n)\mathbf{x}(n) + \mathbf{g}^T(n)\mathbf{x}^*(n)$$

allowing it to model general second-order noncircular signals, that is those, with inequal powers in the real and imaginary parts and this noncircular probability density functions. Such signals are common in practical applications and standard complex valued models are only equipped to deal with circular (proper) signals.

b) As $\mathbb{C}$-derivatives are not defined for real functions of complex variable

$$\mathbb{R}-\text{der:}\quad \frac{\partial}{\partial \mathbf{z}} = \frac{1}{2}\left[\frac{\partial}{\partial \mathbf{x}} - \jmath\frac{\partial}{\partial \mathbf{y}}\right] \qquad \mathbb{R}^*-\text{der:}\quad \frac{\partial}{\partial \mathbf{z}^*} = \frac{1}{2}\left[\frac{\partial}{\partial \mathbf{x}} + \jmath\frac{\partial}{\partial \mathbf{y}}\right]$$

and the gradient

$$\nabla_{\mathbf{w}} J = \frac{\partial J(e,e^*)}{\partial \mathbf{w}} = \left[\frac{\partial J(e,e^*)}{\partial w_1},\ldots,\frac{\partial J(e,e^*)}{\partial w_N}\right]^T = 2\frac{\partial J}{\partial \mathbf{w}^*} = \underbrace{\frac{\partial J}{\partial \mathbf{w}^r} + \jmath\frac{\partial J}{\partial \mathbf{w}^i}}_{pseudogradient}$$

The scalar product

$$< \partial J/\partial \mathbf{w}, \Delta \mathbf{w}^* >= \left[\frac{\partial J}{\partial \mathbf{w}}\right]^H \Delta \mathbf{w}^* = \| \partial J/\partial \mathbf{w} \| \| \Delta \mathbf{w}^* \| \cos\angle(\partial J/\partial \mathbf{w}, \Delta \mathbf{w}^*)$$

achieves its maximum value when the terms in the scalar product are colinear, that is, $\frac{\partial J}{\partial \mathbf{w}} \parallel \Delta \mathbf{w}^*$.

Thus, the maximum change of the gradient of the cost function is in the direction of the conjugate weight vector, and

$$\nabla_{\mathbf{w}} J = \nabla_{\mathbf{w}^*} J \qquad \text{Brandwood}$$

c) **[new example]**
The CMA algorithm is based on the standard minimisation of the mean square (real) error, that is

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu \nabla_{\mathbf{w}} e^2(n)$$

where

$$e(n) = \frac{1}{2}\big(|y(n)|^2 - 1\big)$$

The stochastic gradient setting, the gradient of e^2 becomes

$$\nabla_{\mathbf{w}} e^2(n) = 2e(n)\nabla_{\mathbf{w}}|y(n)|^2$$

Since

$$|y(n)|^2 = \mathbf{w}^H(n)\mathbf{x}(n)\mathbf{x}^H(n)\mathbf{w}(n) \qquad \Rightarrow \qquad \nabla_{\mathbf{w}}|y(n)|^2 = \big[\mathbf{x}(n)\mathbf{x}^H(n)\big]\mathbf{w}(n)$$

and we have

$$\nabla_{\mathbf{w}} e^2(n) = 2e(n)\mathbf{x}(n)\mathbf{x}^H(n)\mathbf{w}(n) = \big(|y(n)|^2 - 1\big)\mathbf{x}(n)\mathbf{x}^H(n)\mathbf{w}(n)$$

thus the CMA algorithm becomes

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\big(1 - |y(n)|^2\big)y^*(n)\mathbf{x}(n)$$

5) a) [**New example**]

$$\begin{aligned} \mathbf{w}(n+1) &= \mathbf{w}(n) - \mu \nabla_{\mathbf{w}} \hat{J}_{|\mathbf{w}=\mathbf{w}(n)} \\ \hat{J} = e^{2p}(n) &\Rightarrow \nabla_{\mathbf{w}} \hat{J} = 2p e^{2p-1}(n) \nabla_{\mathbf{w}} e(n) \\ e(n) &= d(n) - \mathbf{x}^T(n)\mathbf{w}(n) \\ &\Rightarrow \mathbf{w}(n+1) = \mathbf{w}(n) + 2\,p\,\mu e^{2p-1}(n)\mathbf{x}(n) \end{aligned}$$

For p small, the algorithm would still perform satisfactorily on signals with impulsive noise, however for p large, the update would be dominated by the $2p$-th power of the error, and may become very large for large errors. For impulsive artifacts, this would sway the stochastic gradient in the wrong direction, and may cause divergence.
Another issue is the learning rate - the parabolic error performance surface bacomes increasingly narrow with an increase in p, causing the adaptation process to be quite sensitive and to require a very small value of learning rate.

b) [**New example**]
This algorithms operates on the basis of 'mixed norm', combining the 1-norm $E\{|e(k)|\}$ and 2-norm $E\{e^2(k)\}$. Starting from

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

we arrive at the update

$$\begin{aligned} \mathbf{w}(k+1) &= \mathbf{w}(k) + \mu\big[\alpha e(k) + (1-\alpha) sign(e(k))\big]\mathbf{x}(k) \\ &= \mathbf{w}(k) + \underbrace{\mu_1 sign(e(k))\mathbf{x}(k)}_{sign\ LMS} + \underbrace{\mu_2 e(k)\mathbf{x}(k)}_{standard\ LMS} \end{aligned}$$

By varying the convex mixing parameter α, we can move from the purely 'sign' algorithm, based on the minimisation of $|e(k)|$ to the LMS algorithm, based on the minimisation of $e^2(k)$. For impulsive noise with large amplitude, $|e(k)| \ll e^2(k)$ and the sign algorithm is more stable, whereas for smaller error $|e(k)| \gg e^2(k)$ and the LMS has better performance when approaching the steady state. By carefully choosing α or by making it adaptive, we can be benefit from choosing the sign algorithm for large impulsive noise corrupted samples, and LMS otherwise.

c) [**New example**]
Starting from

$$J(k) = \frac{1}{2} e^{e^2(k)} = \frac{1}{2} \sum_{i=0}^{+\infty} \frac{1}{i!} e^{2i}(k)$$

observe that this cost function takes into account allt he even moments of the output error $e(k)$. It therefore generalises the algorithm in a), by proving its 'mixed norm' extension.
For small output errors

$$J(k) \approx \frac{1}{2}\big(1 + e^2(k)\big) \quad \Rightarrow \quad \mathbf{w}(k+1) = \mathbf{w}(k) + \mu e(k)\mathbf{x}(k)$$

10/10

as the constant '1' vanishes when taking the gradients. The algorithm based on an exponential cost function therefore reduces to LMS for small errors. The steepness of the algorithm is, however, very high for large errors, due to the accumulation of high error powers.

d) **new example and intuitive reasoning**

As explained above, the algorithm best suited for situations where large impulsive noise intereference is dominant is the alorithm in b).