Advanced Signal Processing
The Method of Least Squares

Danilo Mandic

room 813, ext: 46271

Department of Electrical and Electronic Engineering
Imperial College London, UK

d.mandic@imperial.ac.uk, URL: www.commsp.ee.ic.ac.uk/∼mandic

Aims

- To introduce the concept of least squares estimation (LSE)
- Parallels with the ML estimation, BLUE, and sample mean estimator
- To introduce signal, noise, and measurement subspaces
- Concept of orthogonality of the signal space and modelling error
- Linear least squares, nonlinear least squares, separable least squares, constrained least squares
- Sequential least squares, link with state space models
- Weighted least squares, confidence levels in data samples
- Practical applications

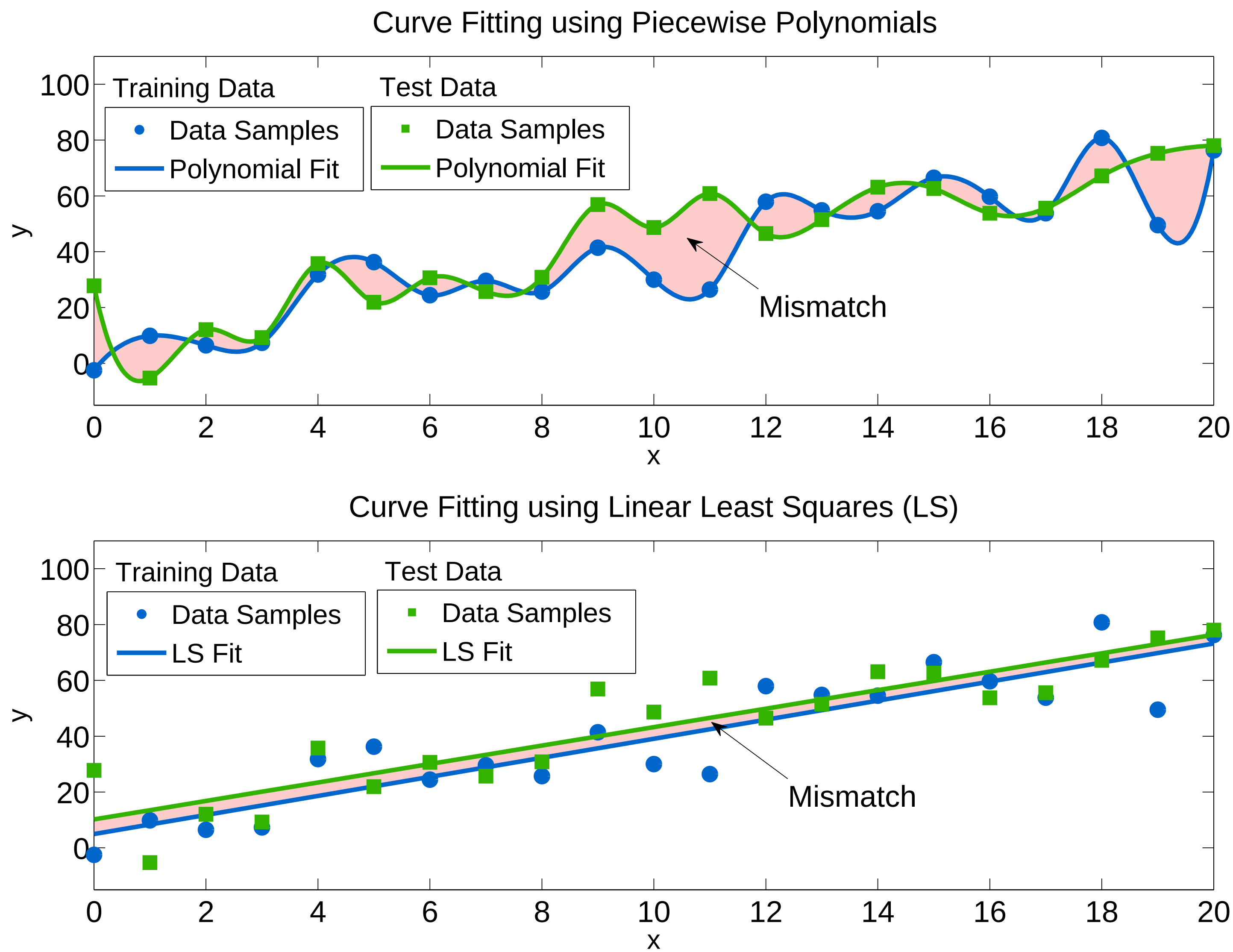
The method of Least Squares

This class of estimators has, generally, no optimality properties

- But, do we necessarily desire optimality $\looparrowright$ an optimal estimator may be mathematically intractable or computationally too complex
- Makes good sense for many practical problems $\looparrowright$ this dates back to Gauss who in 1795 introduced the method to study planetary motions
- No probabilistic assumptions are made about the data, only a signal model is assumed
- Usually easy to implement, either in a block–based or sequential manner, this amounts to the minimisation of a quadratic cost function
- Within the (LS) approach we attempt to minimise the squared difference between the observed data and the assumed model of noiseless data
- Rigorous statistical performance cannot be assessed without some specific assumptions about probabilistic structure in the data

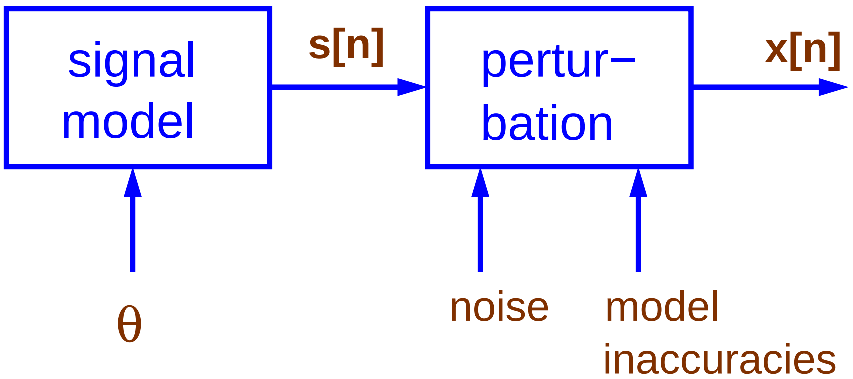
Motivation: A simpler model often generalises better

Consider two models for $x[n] = A + Bn + w[n]$



Data model and the Least Squares Error (LSE) criterion
no probabilistic assumptions made about the data!

The signal $s[n]$ is assumed to be purely deterministic, generated by a model which depends upon an unknown parameter θ or a vector parameter $\boldsymbol{\theta}$.



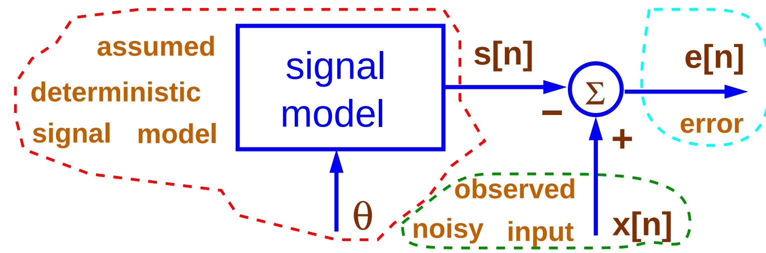
Least squares data model

The observed signal $x[n]$ is subject to:

- ∘ external noise $w[n]$
- ∘ model inaccuracies

No probabilistic assumptions ☺

Only signal model assumed ↬ wide range of applications



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

$$\text{LSE}: \qquad \min_{\theta} J(\theta) \qquad \text{(our objective)}$$

The LS estimator of the unknown parameter θ finds the value of θ that makes the model output $s[n]$ closest to the observed data $x[n]$; the closeness is measured by the LS error criterion (error power)

Example 1: DC Level in WGN

Our old example: DC level in WGN (in MLE, we needed a pdf!)

Data model: $s[n] = A$

Measurement model: $x[n] = s[n] + w[n] = A + w[n], \quad w[n] \looparrowright$ any noise

LSE formulation:

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

LSE solution:

$$\text{set the derivative to zero} \qquad \frac{dJ(A)}{dA} = -2 \sum_{n=0}^{N-1} (x[n] - A) = 0$$

$$\text{the LS estimator}: \qquad \hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

We cannot claim optimality in the MVU sense, except for the Gaussian noise $w \sim \mathcal{N}(0, \sigma^2)$. All we can say is that it the LSE estimator minimises the sum of squared errors (error power).

Still, this leads to a very powerful and practically useful class of estimators.

The method of Least Squares is very convenient
how do we use it in practice?

1. **Problem with signal mean.** If the noise is not zero–mean, then the sample mean estimator actually models $\boldsymbol{x[n] = A + w[n] + w'[n]}$

$$w[n] \sim \text{nonzero mean noise} \quad w'[n] = \text{zero mean noise} \quad \rightarrow \quad E\{x[n]\} = A + E\{w[n]\}$$

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

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

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

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

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

Geometric interpretation & Example: Fourier analysis

Signal model $\mathbf{s} = \mathbf{H}\boldsymbol{\theta} \quad \Leftrightarrow \quad \mathbf{s} = [\underbrace{\mathbf{h}_1, \dots, \mathbf{h}_p}_{\text{columns of } \mathbf{H}}][\theta_1, \dots, \theta_p]^T = \sum_{i=1}^{p} \theta_i \mathbf{h}_i$

$\Rightarrow$ **Signal model is a linear combination of "signal" vectors** $\{\mathbf{h}_1, \dots, \mathbf{h}_p\}$

Example 2: Signal model is $s[n] = a\cos 2\pi f_0 n + b \sin 2\pi f_0 n, \quad f_0$ is known.

Task: Determine the unknown parameters, that is, the amplitudes a, b.

Solution: With f_0 known and $\boldsymbol{\theta} = [a, b]^T$, we have

$$\underbrace{\begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix}}_{\mathbf{s}} = \begin{bmatrix} 1 & 0 \\ \cos 2\pi f_0 & \sin 2\pi f_0 \\ \vdots & \vdots \\ \underbrace{\cos 2\pi f_0[N-1]}_{\mathbf{h}_1} & \underbrace{\sin 2\pi f_0[N-1]}_{\mathbf{h}_2} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \underbrace{[\mathbf{h}_1 \ \mathbf{h}_2]}_{\mathbf{H}} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\boldsymbol{\theta}}$$

$\Rightarrow$ The signal model is

$\mathbf{s} = a\,\mathbf{h}_1 + b\,\mathbf{h}_2 \qquad$ (linear combination of $\mathbf{h}_1$ & $\mathbf{h}_2$); $\qquad$ error $\quad \boldsymbol{\varepsilon} = \mathbf{x} - \mathbf{s}$

and the Least Squares (LS) cost is given by $J(\boldsymbol{\theta}) = \left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right)^T \left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right)$

Geometric interpretation – continued

the signal vector $\mathbf{s}$ is a linear combination of the columns of $\mathbf{H}$

This can be rewritten in a more elegant form.

Recall that the Euclidean length $\|\cdot\|_2$ of an $N \times 1$ vector $\mathbf{q} = [q_1, q_2, \ldots, q_N]^T \in \mathbb{R}^{N\times 1}$ is given by

$$\| \mathbf{q} \|_2 = \sqrt{\sum_{i=1}^{N} q_i^2} = \sqrt{\mathbf{q}^T\mathbf{q}} = \sqrt{< \mathbf{q}, \mathbf{q} >}$$

Then

$$J(\boldsymbol{\theta}) = \left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right)^T \left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right) = \left\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right\|_2^2 = \left\|\mathbf{x} - \sum_{i=1}^{p} \theta_i \mathbf{h}_i\right\|_2^2$$

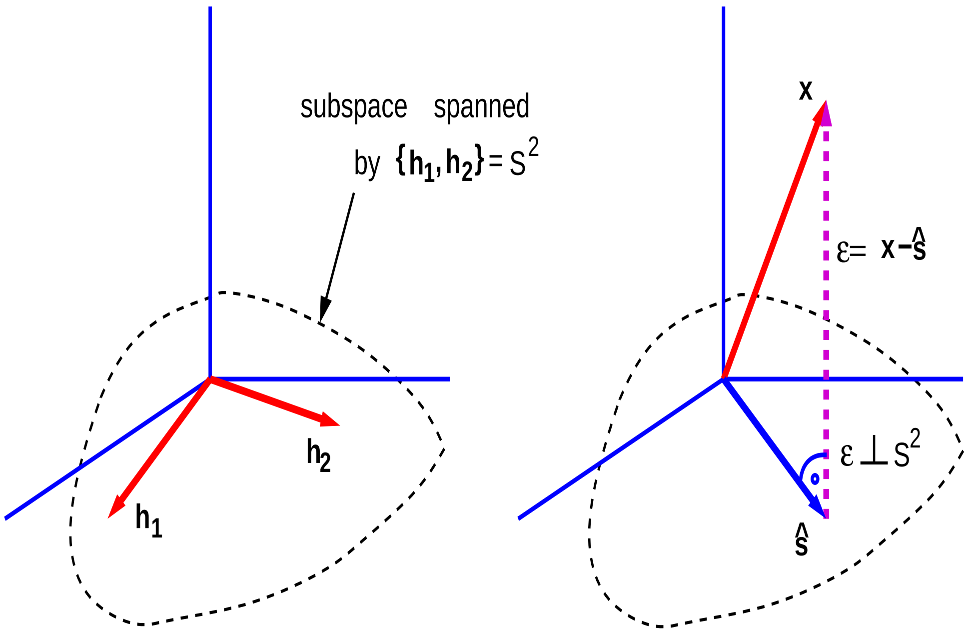
☞ The LSE attempts **to minimise the square of the distance** from the data vector $\mathbf{x}$ to the signal model vector

$$\mathbf{s} = \sum_{i=1}^{p} \theta_i \mathbf{h}_i$$

Vector space projections

signal dimension is lower than measurement dimension (signal lives in a subspace)

The vector $\mathbf{x} \in \mathbb{R}^{N\times 1}$, however, all signal vectors must lie in a p-dimen. 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

⊛ 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, $\hat{\mathbf{s}} = \mathbf{P}\mathbf{x}$, $\mathbf{P}$ ↬ projection matrix.

⊛ 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

$$\boldsymbol{\epsilon} = (\mathbf{x}-\hat{\mathbf{s}}) \perp \mathbf{H} \Leftrightarrow (\mathbf{x}-\mathbf{s}) \perp S^2$$

$$(a): \quad (\mathbf{x}-\mathbf{s}) \perp \mathbf{h}_1 \quad \Rightarrow \quad (\mathbf{x}-\mathbf{s})^T\mathbf{h}_1 = 0$$

$$(b): \quad (\mathbf{x}-\mathbf{s}) \perp \mathbf{h}_2 \quad \Rightarrow \quad (\mathbf{x}-\mathbf{s})^T\mathbf{h}_2 = 0$$

Finally: Least squares solution

Using

$$\mathbf{s} = \theta_1 \mathbf{h_1} + \theta_2 \mathbf{h_2} = \mathbf{H}\boldsymbol{\theta}$$

and from the conditions (a) and (b), we have

$$\begin{aligned} \left(\mathbf{x} - \theta_1 \mathbf{h_1} - \theta_2 \mathbf{h_2}\right)^{\mathbf{T}} \mathbf{h_1} &= 0 \qquad \equiv \qquad \boldsymbol{\varepsilon}^T \mathbf{h}_1 = 0 \\ \left(\mathbf{x} - \theta_1 \mathbf{h_1} - \theta_2 \mathbf{h_2}\right)^{\mathbf{T}} \mathbf{h_2} &= 0 \qquad \equiv \qquad \boldsymbol{\varepsilon}^T \mathbf{h}_2 = 0 \end{aligned}$$

Since $\mathbf{H} = [\mathbf{h}_1 \ \mathbf{h}_2]$ and $\boldsymbol{\theta} = [a \ b]^T$, these conditions can be combined into a vector/matrix form

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

Now, use the identity $(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^{\mathbf{T}}\mathbf{H} = \mathbf{H}^{\mathbf{T}}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$ and solve for the unknown vector param. $\boldsymbol{\theta}$, to yield the **Least Squares Estimator (LSE)**

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

where $\mathbf{H}$ is the $(N \times p)$-dimensional measurement (observation) matrix.

Example 2: Fourier analysis $\looparrowright$ continued

for more detail see Example 9 in Lecture 4

For $f_0 = k/N$, where $k = 1, 2, \ldots, N/2 - 1$, then the scalar product of the columns of the observation matrix $\mathbf{H}$ becomes (orthogonality etc)

$$\mathbf{h}_1^T \mathbf{h}_2 \quad = \quad \sum_{n=0}^{N-1} \cos\left(2\pi \frac{k}{N} n\right) \sin\left(2\pi \frac{k}{N} n\right) = 0 \quad \Leftrightarrow \quad \mathbf{h}_1 \perp \mathbf{h}_2 \quad \text{(orthogonal)}$$

$$\text{while} \qquad \mathbf{h}_1^T \mathbf{h}_1 = \frac{N}{2} \qquad \mathbf{h}_2^T \mathbf{h}_2 = \frac{N}{2}$$

which means that $\mathbf{h}_1$ and $\mathbf{h}_2$ are orthogonal but not orthonormal.

Combining the above results gives $\mathbf{H}^T \mathbf{H} = \frac{N}{2} \mathbf{I}$ and therefore

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{x} = \frac{2}{N} \mathbf{H}^T \mathbf{x} = \begin{bmatrix} \frac{2}{N} \sum_{n=0}^{N-1} x[n] \cos(2\pi \frac{k}{N} n) \\ \frac{2}{N} \sum_{n=0}^{N-1} x[n] \sin(2\pi \frac{k}{N} n) \end{bmatrix}$$

In general the columns of $\mathbf{H}$ will not be orthogonal, so that the signal vector estimate is obtained as

$$\hat{\mathbf{s}} = \mathbf{H} \hat{\boldsymbol{\theta}} = \underbrace{\mathbf{H} \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T}_{\text{projection matrix } \mathbf{P}} \mathbf{x} = \mathbf{P} \mathbf{x}$$

Linear least squares in a nutshell

Suppose a linear observation model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. Then the **cost function**

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{n=0}^{N-1} \big(x[n] - s[n,\theta]\big)^2 = \big(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\big)^T \big(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\big) \\ &= \mathbf{x}\mathbf{x}^T - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} \qquad \mathbf{H} \text{ is full rank} \end{aligned}$$

The gradient of the cost function is then

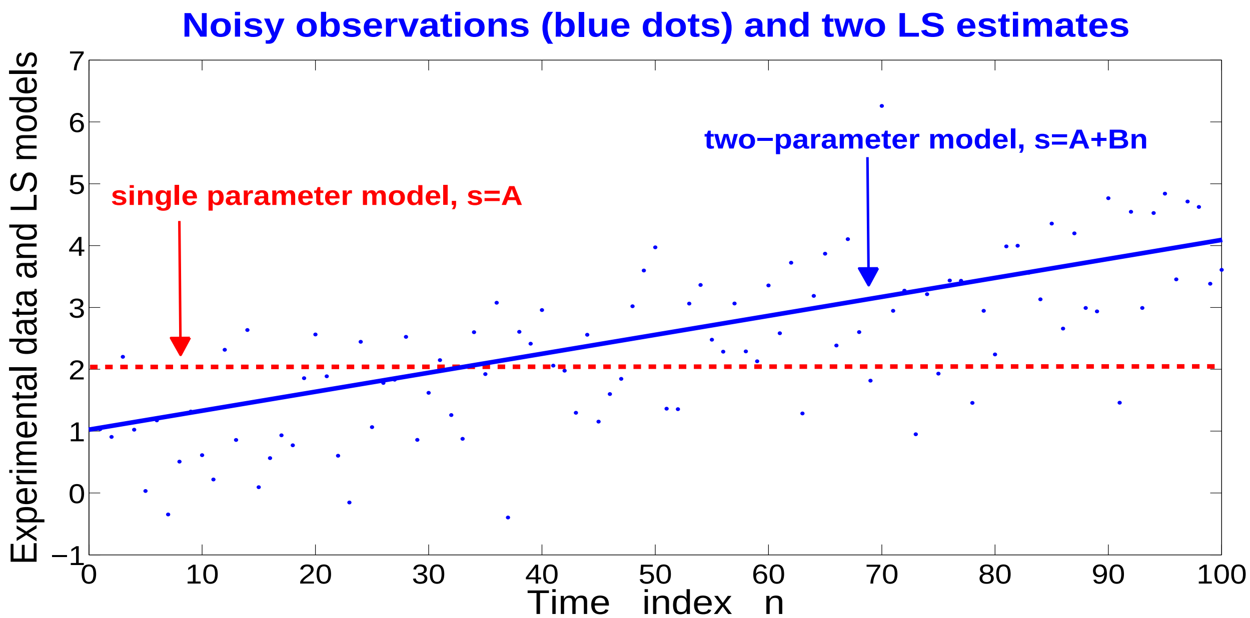
$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} = \mathbf{0}$$

1. The LSE estimator $\qquad \hat{\boldsymbol{\theta}} = \big(\mathbf{H}^T\mathbf{H}\big)^{-1}\mathbf{H}^T\mathbf{x}$
2. The minimum LS cost (replace $\hat{\boldsymbol{\theta}}$ into $J(\boldsymbol{\theta})$ above) is therefore

$$J_{min} = J(\hat{\boldsymbol{\theta}}) = \mathbf{x}^T\big[\mathbf{I} - \mathbf{H}\big(\mathbf{H}^T\mathbf{H}\big)^{-1}\mathbf{H}^T\big]\mathbf{x} = \mathbf{x}^T\big(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\big)$$

Linear least squares in a nutshell, continued

- The LS approach can be interpreted as the problem of approximating a data vector $\mathbf{x} \in \mathbb{R}^N$ by another vector $\hat{\mathbf{s}}$ which is a linear combination of vectors $\{\mathbf{h}_1, \ldots, \mathbf{h}_p\}$ that lie in a p-dimensional subspace $S \in \mathbb{R}^p \in \mathbb{R}^N$
- The problem is solved by choosing $\hat{\mathbf{s}}$ so as to be an orthogonal projection of $\mathbf{x}$ on the subspace spanned by $\mathbf{h}_i, i = 1, \ldots, p$
- The LS estimator is very sensitive to the correct deterministic model of $\mathbf{s}$, as shown in the figure below for the LS fit of $x[n] = A + Bn + w[n]$.



The role of the model order p

think the order of an AR process $x(n) = a_1 x(n-1) + \cdots + a_p x(n-p) + w(n)$

Follows naturally from the problem of fitting a polynomial to the data (recall the Weierstrass theorem - any continuous differentiable function can be approximated arbitrarily well with a high-enough order polynomial)

- Observe that J_{min} is a **non-increasing function** of the model order p
- The choice $p = N$ is a perfect fit for the data, but this way we also fit the noise (see also Slide 4)
- Recall the MDL and AIC in AR modelling - we choose the **simplest model order** p that is adequate for the data
- If we have a specified J_{min} then we can gradually increase p until we reach the required J_{min}
- **To save on computation, we can also use an order-recursive LS algorithm to compute the model of order $(p+1)$ from the model of order p**

Weighted Least Squares (WLS)

see also Example 5 in Lecture 5

To emphasize the contribution of those data samples that are deemed to be more reliable, we can include an $N \times N$ positive definite (and hence symmetric) weighting matrix $\mathbf{W}$ so that

$$J(\boldsymbol{\theta}) = \big(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\big)^T \mathbf{W} \big(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\big)$$

It is straightforward to show that the weighted least squares solution

$$\hat{\boldsymbol{\theta}} = \big(\mathbf{H}^T\mathbf{W}\mathbf{H}\big)^{-1}\mathbf{H}^T\mathbf{W}\mathbf{x} \quad \& \quad J_{min} = \mathbf{x}^T\Big(\mathbf{W} - \mathbf{W}\mathbf{H}\big(\mathbf{H}^T\mathbf{W}\mathbf{H}\big)^{-1}\mathbf{H}^T\mathbf{W}\Big)\mathbf{x}$$

Example 3: For a diagonal $\mathbf{W}$ with elements $[\mathbf{W}]_{ii} = w_i > 0$, the LS error of the DC level estimator becomes

$$J(A) = \sum_{n=0}^{N-1} w_n \big(x[n] - A\big)^2$$

If $x[n] = A + q[n]$, where the zero-mean **uncorrelated** noise (of any distribution) $q[n] \sim (0, \sigma_n^2)$, it is reasonable to choose $w_n = 1/\sigma_n^2$, to give

$$\hat{A} = \Big(\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}\Big)\Big(\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}\Big)^{-1}$$

Remark: If we take $\mathbf{W} = \mathbf{C}^{-1}$, then we have the BLUE estimator.

Opportunities in practical applications $\looparrowright$ numerous

- **Constrained least squares.** We can incorporate a set of linear constraints in the form $\mathbf{A}\boldsymbol{\theta} = \mathbf{c}$, to have a constrained LS criterion
$$J_c(\boldsymbol{\theta}) = \big(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\big)^T\big(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\big) - \boldsymbol{\lambda}\big(\mathbf{A}\boldsymbol{\theta} - \mathbf{c}\big)$$
Use e.g. Lagrange optimisation to solve (first term $\looparrowright$ LS solution $\hat{\boldsymbol{\theta}}$).
- **Nonlinear least squares.** The signal model is nonlinear, i.e. $\mathbf{s} \neq \mathbf{H}\boldsymbol{\theta}$
We can either linearise the problem (e.g. using Taylor series expansion) or solve it numerically in some iterative or recursive fashion. These methods are often prone to convergence problems if highly nonlinear.
- **Dealing with nonlinear least squares - parameter transformation.**
Example: Consider a nonlinear problem of estimating the amplitude and phase of a sinusoid $\quad s[n] = A\cos(\omega n + \phi), \quad n = 0, \ldots, N-1$

$\rightsquigarrow$ Transform the problem into $\boldsymbol{A\cos(\omega n + \phi) = A\cos\phi\cos\omega n - A\sin\phi\sin\omega n}$

Variable swap. Let $\alpha_1 = A\cos\phi$ and $\alpha_2 = -A\sin\phi$, and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^T$.
Now, the signal model becomes linear in $\boldsymbol{\alpha}$, that is, $\mathbf{s} = \mathbf{H}\boldsymbol{\alpha}$

Use LS to obtain $\hat{\boldsymbol{\alpha}} = \big(\mathbf{H}^T\mathbf{H}\big)^{-1}\mathbf{H}^T\mathbf{x}$

where $A = \sqrt{\alpha_1^2 + \alpha_2^2}$ and $\phi = \arctan(-\alpha_2/\alpha_1)$

LS estimation in the big picture of estimators

Consider the linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$

Estimator	Model	Assumption	Estimate
LSE	$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$	no probabilistic assumptions	$\hat{\boldsymbol{\theta}}_{ls} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$
BLUE	$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$	w is white with unknown pdf	$\hat{\boldsymbol{\theta}}_{blue} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$
MLE	$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$	need to know pdf of w	$\hat{\boldsymbol{\theta}}_{mle} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$
MVUE	$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$	need to know pdf of w	$\hat{\boldsymbol{\theta}}_{mvu} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$

LSE and orthogonal projections:

Signal model is $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$ $\looparrowright$ the estimate is a projection of $\mathbf{x}$ onto $S^p \in \mathbb{R}^p \subset \mathbb{R}^N$

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

where $\mathbf{P} = \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ is called the **projection matrix**. Since the estimated signal $\hat{\mathbf{s}} = \mathbf{P}\mathbf{x} \in S^p$, it follows that $\mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}\mathbf{x}$.

Therefore, any projection matrix is **idempotent**, that is $\mathbf{P}^2 = \mathbf{P}$, it is symmetric and **singular** with rank p (many x(n) can have the same projection).

Sequential least squares

Oftentimes in signal processing, data are being collected sequentially, namely one point at a time. To process such data, we can either:

- wait until all the data points (samples) are collected and make an estimate of the unknown parameter $\looparrowright$ **block-based approach**, or
- refine our estimate as each new sample arrives $\looparrowright$ **sequential approach**

We therefore need to obtain a sequence of LS estimators over time.

The problem:

Given a known least squares estimate, $\hat{\boldsymbol{\theta}}_{N-1}$, which is based on the signal history (all the data samples in the past)

$$\{x[0], x[1], \ldots, x[N-1]\}$$

we need to produce a new estimate, $\hat{\boldsymbol{\theta}}_{N}$, upon observing the new available data sample $x[N]$.

Question: Can we update the existing solution $\hat{\boldsymbol{\theta}}_{N-1}$ sequentially, without having to solve the LS equations again from scratch?

Example 4: DC level in uncorrelated zero mean noise

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

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

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

$$\begin{aligned}\hat{A}[N] &= \frac{1}{N+1}\sum_{n=0}^{N} x[n] = \frac{1}{N+1}\Big(\sum_{n=0}^{N-1} x[n] + x[N]\Big)\\ \hat{A}[N] &= \frac{N}{N+1}\hat{A}[N-1] + \frac{1}{N+1}x[N] \qquad \hookrightarrow \ \textbf{a recursive estimate!}\end{aligned}$$

The minimum LS error can also be computed recursively, as

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

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

Example 4: DC level in noise $\looparrowright$ a more convenient form of the sequential estimator and the associated MSE

Clearly, the new estimate $\hat{A}[N]$ can be calculated from the old estimate $\hat{A}[N-1]$, upon receiving the new observation $x[N]$.

The solution can be rewritten in a more physically insightful form, as

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

The minimum LS error then becomes (show yourselves using (*))

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

Notice that $J_{\min}$ increases with the number of data points N, as we are trying to fit more points with the same number of parameters.

Example 5: Weighted LS for the estimation of a DC level in noise in the sequential form (see also Example 9 in Lecture 4)

Start from

$$J(A) = \sum_{n=0}^{N-1} w_n \big(x[n] - A\big)^2$$

If $x[n] = A + q[n]$, where the zero-mean **uncorrelated** noise (any distribution) $q[n] \sim (0, \sigma_n^2)$, it is reasonable choose $w_n = 1/\sigma_n^2$, to give[1]

$$\textbf{Standard LS solution :} \qquad \hat{A}[N] = \frac{\sum_{n=0}^{N} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}}$$

Its corresponding sequential form then becomes

$$\hat{A}[N] = \hat{A}[N-1] + \frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} \big(x[N] - \hat{A}[N-1]\big)$$

and has the form

new estimate $=$ old estimate $+$ gain $\times$ error

[1] In standard weighted LS, with a diagonal weighting matrix $\mathbf{W}$ we would have $[\mathbf{W}]_{ii} = \frac{1}{\sigma_i^2}$

Some observations about weighted LS

Notice that the gain factor that multiplies the correction term now **depends on our confidence** in the new data sample, given by $1/\sigma_N^2$.

- If $\sigma_N^2 \rightarrow \infty$, i.e. the new sample is noisy, we do not correct the previous LSE

- If $\sigma_N^2 \rightarrow 0$, that is, the new sample is noise–free, then $\hat{A} \rightarrow x[N]$, and we discard all the previous samples

☞ If we assume $x[n] = A + w[n]$, with $\{w[n]\}$ zero mean uncorrelated noise for which the variance of each $w[n]$ is $\sigma_n^2,\ n = 0, \ldots, N-1$, then the LSE is also the BLUE and

$$var\big(\hat{A}[N-1]\big) = \frac{1}{\displaystyle\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}$$

Weighted LS: Recursive calculation of gain and variance

- The gain factor for the N-th correction can be written as

$$K[N] = \frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} = \frac{var\big(\hat{A}[N-1]\big)}{var\big(\hat{A}[N-1]\big) + \sigma_N^2}$$

- Since $0 < K[N] \leq 1$, the correction in a sequential estimator is large if $K[N]$ is large or $var\big(\hat{A}[N-1]\big)$ is large

- Similarly, if the variance of the previous estimate is small, then so too is the correction

- The recursive expression for the variance can be calculated as

$$var\big(\hat{A}[N]\big) = \Big(1 - K[N] var\big(\hat{A}[N-1]\big)\Big)$$

Notice that the gain $K[n]$ is also a random variable.

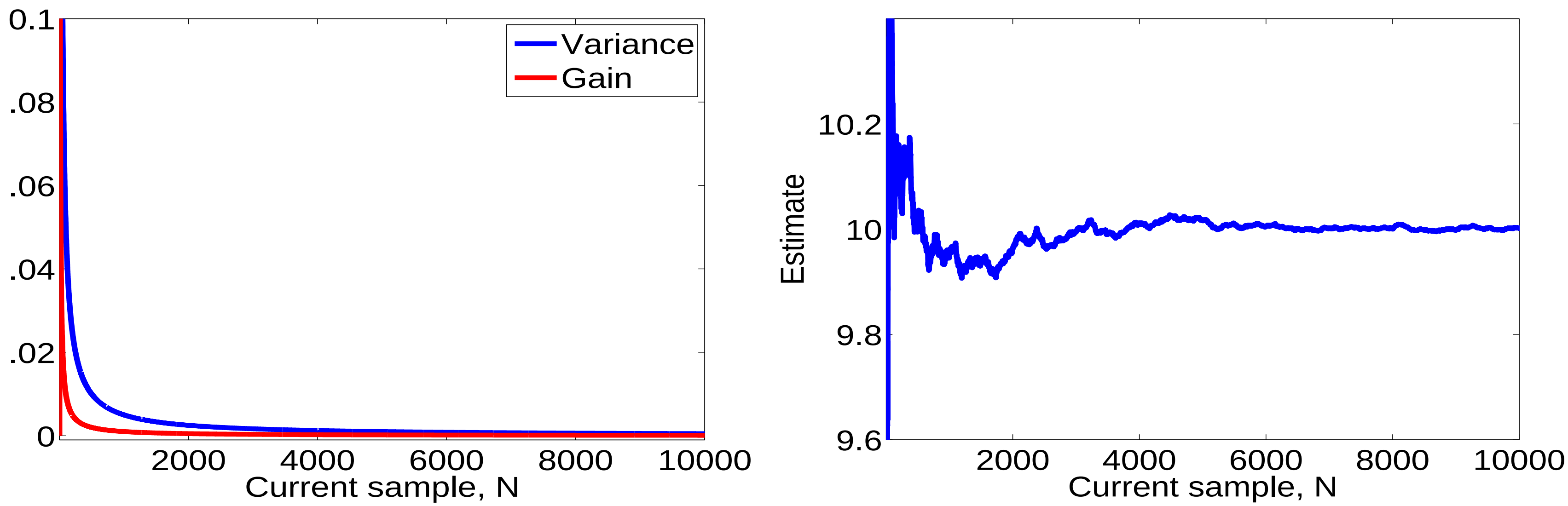
Example 5: Summary of sequential DC level estimators (both weighted and standard)

Estimator update: $$\hat{A}[N] = \hat{A}[N-1] + K[N]\Big(x[N] - \hat{A}[N-1]\Big)$$

$$\text{where} \quad K[N] = \frac{var\big(\hat{A}[N-1]\big)}{var\big(\hat{A}[N-1]\big) + \sigma_N^2}$$

Variance update: $$var\big(\hat{A}[N]\big) = (1 - K[N])\,var\big(\hat{A}[N-1]\big)$$

Initialisation: $$\hat{A}[0] = x[0], \quad var\big(\hat{A}[0]\big) = \sigma_0^2$$



Sequential DC level estimation: $A = 10, \sigma_w^2 = 5$. **Left: Variance and gain. Right: The estimate.**

Sequential LSE for a vector parameter

For a data vector $\mathbf{x}[n] = [x[0], x[1], \ldots, x[n]]^T \rightsquigarrow \quad \mathbf{H}[n] = \begin{bmatrix} \mathbf{H}[n-1]_{n \times p} \\ \\ \mathbf{h}^T[n]_{1 \times p} \end{bmatrix}$

Note that the size of the observation matrix $\mathbb{H}$ grows with time.

- **Estimator update:**

$$\hat{\boldsymbol{\theta}}[n] = \hat{\boldsymbol{\theta}}[n-1] + \mathbf{K}[n]\Big(\mathbf{x}[n] - \mathbf{h}^T[n]\boldsymbol{\theta}[n-1]\Big)$$

where the **gain factor** is given by

$$\mathbf{K}[n] = \mathbf{C}[n-1]\mathbf{h}[n]\left[\sigma_n^2 + \mathbf{h}^T[n]\mathbf{C}[n-1]\mathbf{h}[n]\right]^{-1}$$

- **Covariance matrix update:**

$$\mathbf{C}[n] = \Big(\mathbf{I} - \mathbf{K}[n]\mathbf{h}^T[n]\Big)\mathbf{C}[n-1]$$

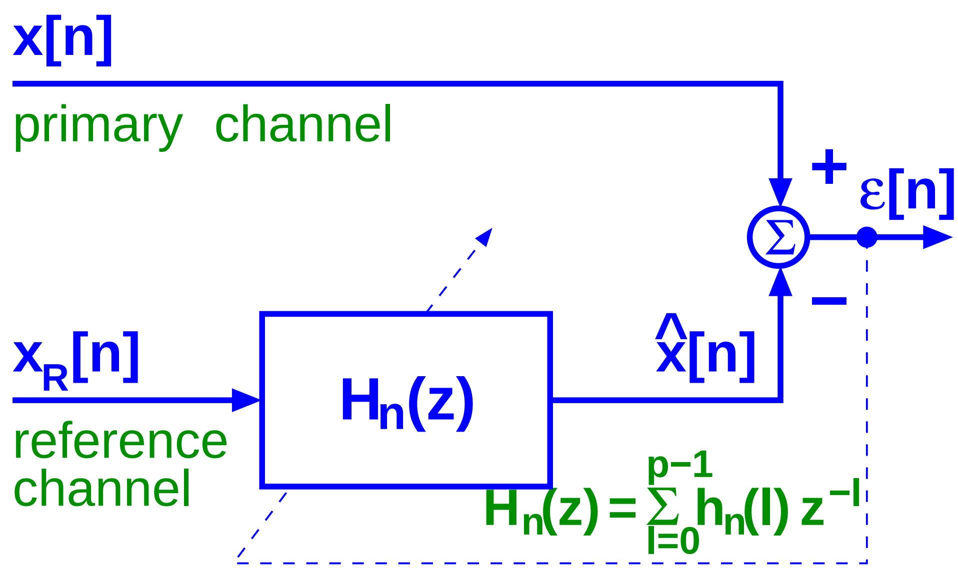
- **Initialisation:** $\mathbf{C}[-1] = \alpha\mathbf{I}, \quad \alpha\text{-large}, \qquad \boldsymbol{\theta}[-1] = \mathbf{0}$

Case study: Adaptive Noise Canceller (ANC)

common paradigm in signal processing is to reduce unwanted noise

Example 1: we may wish to remove background noise in aircraft and car audio systems (noise cancelling headphones, road noise cancellation)

Example 2: a common problem is 50Hz mains artefact in biomedical instrumentation



The configuration of a sequential noise canceller

The reference channel takes the role of the traditional input and the primary channel is the noisy signal of interest.

ANC $\looparrowright$ line interference removal

- **Primary channel:** 'signal' + 'noise to be cancelled' (50 Hz interference)
- **Reference channel:** noise source which is related to the noise in the primary channel (nonzero correlation)
- Filter coefficients are updated sequentially to make $\hat{x}[n]$ as close to $x[n]$ as possible, in the LS sense
- We therefore desire to ensure $\varepsilon[n] = 0$, by minimising the error power

$$\begin{aligned} J[n] &= \sum_{k=0}^{n} \varepsilon^2[k] = \sum_{k=0}^{n} \big(x[k] - \hat{x}[k]\big)^2 \\ &= \sum_{k=0}^{n} \Big(x[k] - \sum_{l=0}^{p-1} h_n(l) x_R[k-l]\Big)^2 \end{aligned}$$

- Filter coefficients (weights) can then be determined as a solution of the sequential LS problem

ANC $\looparrowright$ some practical considerations

The signal and noise are typically statistically nonstationary, and to deal with that we introduce a **weighting or "forgetting factor"** λ, for which the range $0 < \lambda < 1$, so that the cost function becomes

$$
\begin{aligned}
J[n] &= \sum_{k=0}^{n} \lambda^{n-k} \Big(x[k] - \sum_{l=0}^{p-1} h_n(l) x_R[k-l] \Big)^2 \\
&\text{or} \\
J'[n] &= J[n]\lambda^{-n} = \sum_{k=0}^{n} \frac{1}{\lambda^k} \Big(x[k] - \sum_{l=0}^{p-1} h_n(l) x_R[k-l] \Big)^2
\end{aligned}
$$

☞ This is also the form of the standard weighted LS problem.

The sequential LS vector estimator of the filter coefficients is denoted by

$$
\hat{\boldsymbol{\theta}}[n] = \big[\hat{h}_n(0), \hat{h}_n(1), \ldots, \hat{h}_n(p-1) \big]^T
$$

ANC summary

Input reference vector: $\mathbf{x}[n] = \big[x_R[n], x_R[n-1], \ldots, x_R[n-p+1]\big]^T$

Weights: $\sigma_n^2 = \lambda^n$ weighting coefficients w ☞ forgetting factor λ

Error: $e[n] = x[n] - \mathbf{x}^T[n]\hat{\boldsymbol{\theta}}[n-1] = x[n] - \sum_{l=0}^{p-1} \hat{h}_{n-1}(l) x_R[n-l]$

Estimator update: $\hat{\boldsymbol{\theta}}[n] = \hat{\boldsymbol{\theta}}[n-1] + \mathbf{K}[n]e[n]$

where

$$
\begin{aligned}
e[n] &= x[n] - \sum_{l=0}^{p-1} \hat{h}_{n-1}(l) x_R[n-l] \\
\mathbf{K}[n] &= \frac{\mathbf{C}[n-1]\mathbf{x}[n]}{\lambda^n + \mathbf{x}^T[n]\mathbf{C}[n-1]\mathbf{x}[n]} \\
\mathbf{x}[n] &= \big[x_R[n], x_R[n-1], \ldots, x_R[n-p+1]\big]^T \\
\mathbf{C}[n] &= \big(\mathbf{I} - \mathbf{K}[n]\mathbf{x}^T[n]\big)\mathbf{C}[n-1], \quad \text{typically} \quad 0.9 < \lambda < 1
\end{aligned}
$$

In LS methods we do not know the probability densities or σ_n^2 for every sample $x[n]$.

☞ we replace them with a forgetting factor λ^n. This favours most recent samples 👍

Example 6: ANC for line noise removal (0.1Hz sinus. interfer.)

reference x_R is correlated with interference but has different amplitude and phase

○ Interference $x[n] = 10\cos(2\pi(0.1)n + \pi/4)$

○ The reference noise: $x_R[n] = \cos(2\pi(0.1)n)$

○ **Initialisation:** $\hat{\boldsymbol{\theta}}[-1] = \mathbf{0}$, $\mathbf{C}[-1] = 10^5\mathbf{I}$, and $\lambda = 0.99$

○ We need two filter coefficients to model the amplitude and phase of the interference, that is

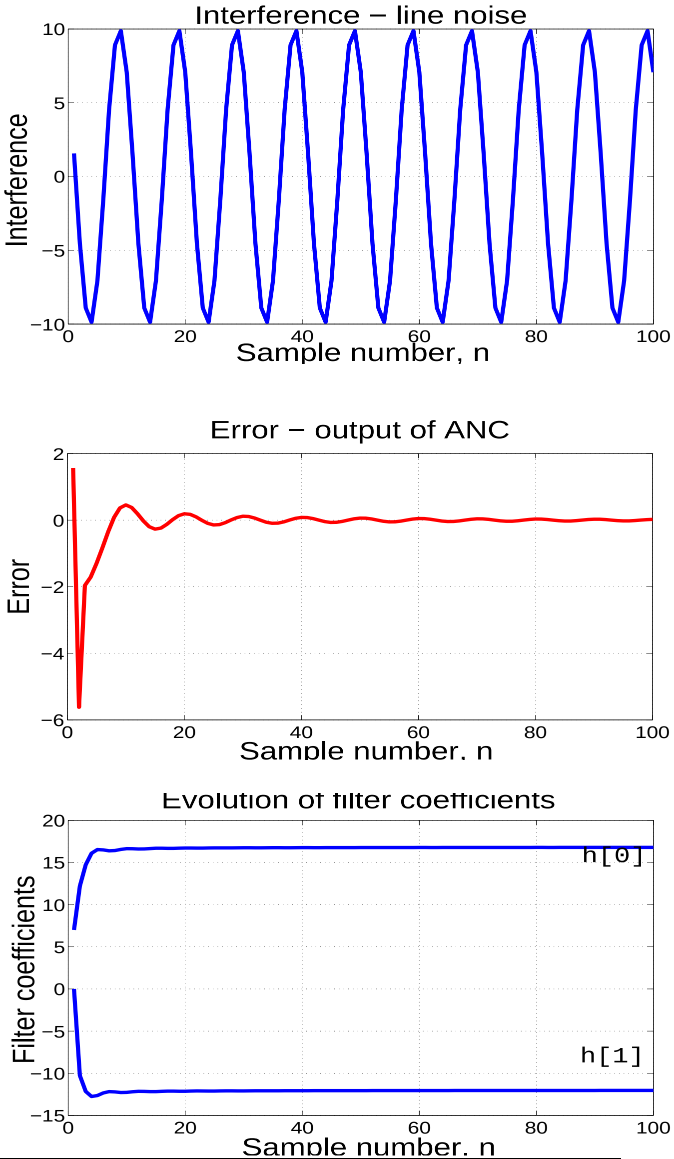
$$\mathcal{H}[exp(2\pi(0.1))] = 10exp(\jmath\pi/4)$$

$\rightsquigarrow$ the adaptive filter must increase the gain of the reference by 10 and phase by $\pi/4$ to match the interference.

Solving, we have

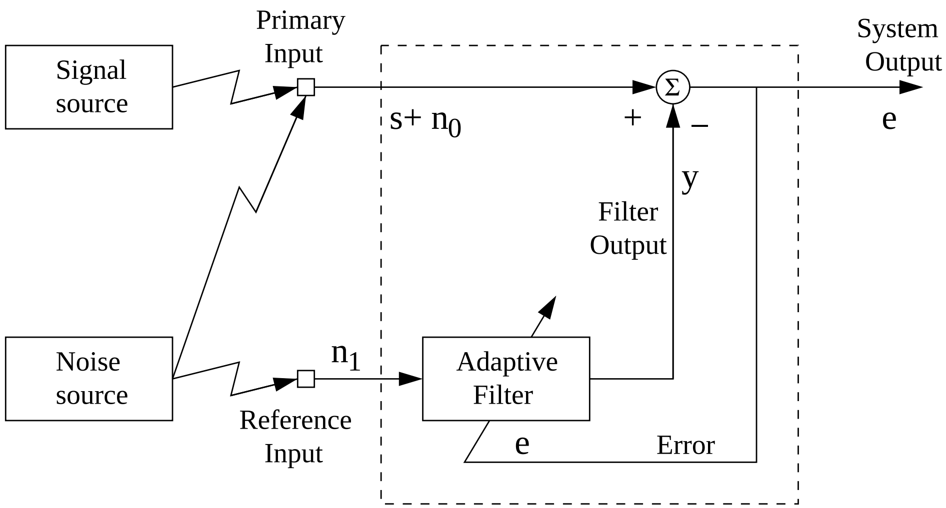
$h[0] + h[1]exp(-2\jmath\pi(0.1)) = 10exp(\jmath\pi/4)$

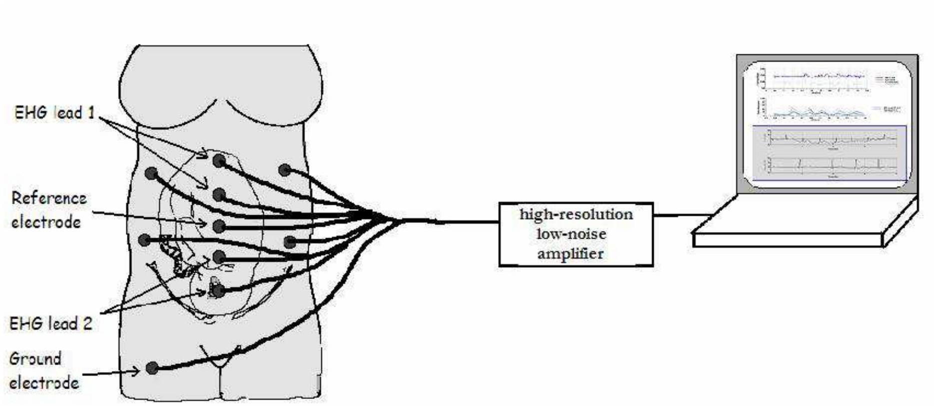
which results in $h[0] = 18.8$ and $h[1] = -12$.



Example 7: Foetal ECG recovery
Data acquisition

ANC with reference



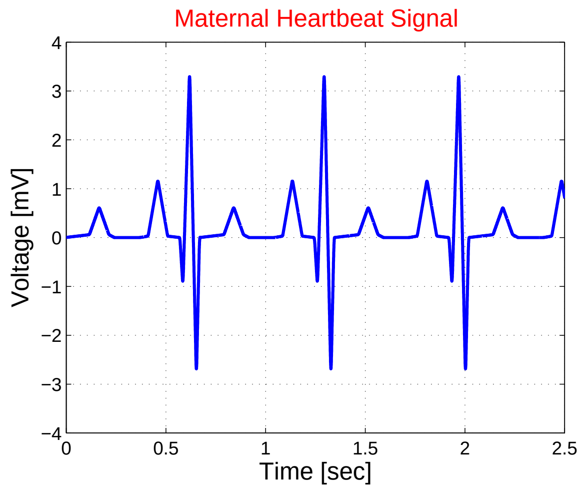


ECG recording (Reference electrode $\neq$ Reference input)

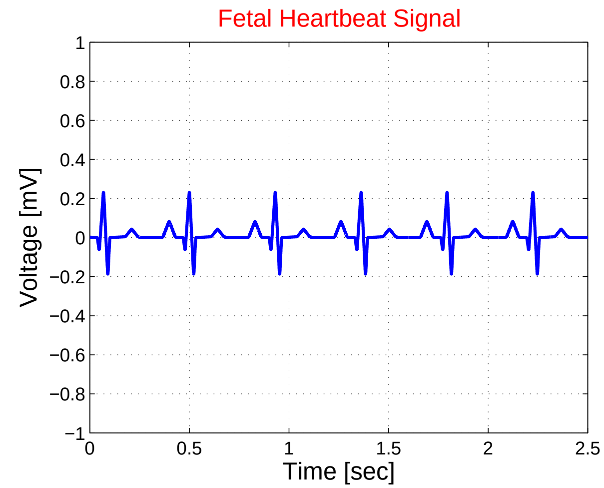
Example 7: Foetal ECG (using slightly updated method)

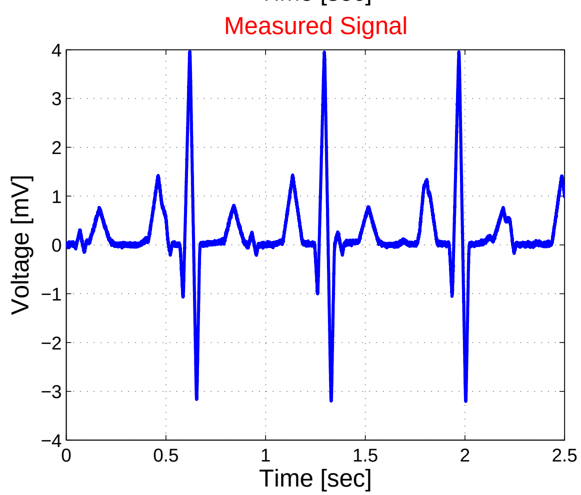
we measure foetal ECG from mother's tammy $\Rightarrow$ ECG(mother) + ECG(baby)

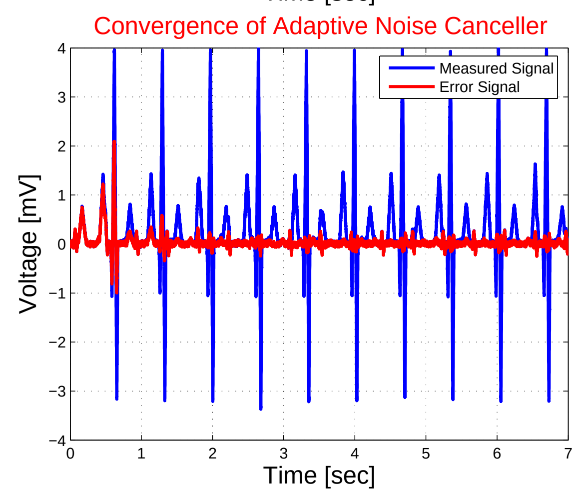
Maternal ECG signal



Foetal heartbeat







Measured ECG (mother ECG + baby ECG)

Error convergence

Lecture summary

- The method of least squares is extremely important for practical applications
- No assumption on the PDF or any other statistics needed
- Estimation error orthogonal to the signal model space
- If the signal model (which is deterministic) is inaccurate, the LS estimator will be biased
- Easy to implement and straightforward to interpret
- Sequential solutions to the LS problem are very practical
- Weighted least squares allow to assign "confidence" to samples
- We can also use a forgetting factor to deal with time-varying statistics
- A number of applications of LS theory: adaptive noise cancellation, digital filter design, Prony type spectral estimation, and many more

Appendix: Derivation of the MMSE and variance for the sequential estimator of a DC level in noise

$$
\begin{aligned}
J_{min}[N] &= \sum_{n=0}^{N}\left(x[n]-\hat{A}[N]\right)^2 \qquad J_{min}[N-1] = \sum_{n=0}^{N-1}\left(x[n]-\hat{A}[N-1]\right)^2 \\
&= \sum_{n=0}^{N-1}\Big[x[n]-\hat{A}[N-1]-\frac{1}{N+1}\left(x[N]-\hat{A}[N-1]\right)\Big]^2 + \left(x[N]-\hat{A}[N]\right)^2 \\
&= J_{min}[N-1]-\frac{2}{N+1}\sum_{n=0}^{N-1}\left(x[n]-\hat{A}[N-1]\right)\left(x[N]-\hat{A}[N-1]\right) \\
&\quad + \frac{N}{(N+1)^2}\left(x[N]-\hat{A}[N-1]\right)^2 + \left(x[N]-\hat{A}[N]\right)^2 \\
J_{min}[N] &= J_{min}[N-1]+\frac{N}{N+1}\left(x[N]-\hat{A}[N-1]\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}\left(\hat{A}[N]\right) &= \frac{1}{\sum_{n=0}^{N}\frac{1}{\sigma_n^2}} = \frac{1}{\sum_{n=0}^{N-1}\frac{1}{\sigma_n^2}+\frac{1}{\sigma_N^2}} = \frac{1}{\frac{1}{\mathrm{var}(\hat{A}[N-1])}+\frac{1}{\sigma_N^2}} \\
&= \frac{\mathrm{var}(\hat{A}[N-1])\ \sigma_N^2}{\mathrm{var}(\hat{A}[N-1])+\sigma_N^2} = \left(1-\frac{\mathrm{var}(\hat{A}[N-1])}{\mathrm{var}(\hat{A}[N-1])+\sigma_N^2}\right)\mathrm{var}(\hat{A}[N-1]) \\
&= \left(1-K[N]\right)\mathrm{var}(\hat{A}[N-1])
\end{aligned}
$$

Appendix: Derivation of the MMSE and variance for the sequential estimator of a DC level in noise

Notes:

○

Notes:

○

Notes:

○

Notes:

-