Lecture 16: Method of Least Squares

- Can be seen as a generalization of the Wiener filter theory.
- Wiener filter introduces assumptions of stationarity.
- Partial application of Wiener filter usually requires periodicity of time series.
- LS-deterministic approach of "fitting" the data - no underlying assumptions on data statistics.
- Data fitting is one of the oldest and the most important problems in engineering.

Statement of the Linear Least-Squares Estimation Problem

\[ u_i \rightarrow \text{physical system} \rightarrow d_i \]

- Physical system performs mapping between \( u_i \rightarrow d_i \).
- Task of model development is to develop a mathematical set of equations that approximates the mapping performed by the system.

Model development usually consists of two steps:

1) Postulate the architecture of the model - choose a set of adjustable parameters.
2) Use pairs \((u_i, d_i), i = 1, N\) to determine the values of adjustable parameters so that they optimize some suitable chosen cost function.
   - by far, the most popular cost function is the sum of squares of model prediction errors \( \Rightarrow \text{LMS estimate} \)

For linear LMS estimation, we use

- Linear Combiner as a model for system architecture
- MSE (Mean Square Error) as the cost function
Linear regression equation (model architecture)

\[ d_i = \omega^H \cdot u_i + e_i, \quad i = 1, \ldots, N \]

- \( d_i \): desired output
- \( u_i \): input vector (assumed as zero mean)
- \( \omega \): model parameters
- \( e_i \): error of the model

\[ J(\omega) = \frac{1}{N} \sum_{i=1}^{N} |e_i|^2 = \frac{1}{N} \sum_{i=1}^{N} e_i e_i^* \approx \sum_{i=1}^{N} e_i e_i^* \]

Define

\[ d = [d_1, d_2, \ldots, d_N]^H, \quad e^H = [e_1, e_2, \ldots, e_N] \]

\[ A = [u_1, u_2, \ldots, u_N]^H \] - data matrix
Then we have

\[ d_1^* = u_1^w + e_1^* \]
\[ d_2^y = u_2^y + e_2^* \]
\[ d^x = u^w + e_3^* \]

\[ \tilde{d} = A \cdot \omega + \varepsilon \] - walm x statement of (*)

The cost function can be expressed as.

\[ J(\omega) = e^H e = \sum (e_i^* e_i) = \frac{1}{2} \| \varepsilon \|^2 = H x E \]

From walm x formula for the model

\[ e = d - A \omega \]

\[ J(\omega) = e^H e = (d - A \omega)^H (d - A \omega) \]

\[ = d^H d - w^H A^H d - d^H A \omega + \omega^H A^H A \omega \]

Minimizing \( J(\omega) \)

\[ \frac{\partial J(\omega)}{\omega} = 0 - 2 A^H d + 2 A^H A \omega = 0 \]

or

\[ A^H A \omega = A^H d \]

If \( A^H A \) is invertible (i.e., nonsingular)
The optimal weights are given as

$$w_{opt} = (A^H A)^{-1} A^H d$$

The matrix $A^H A$ is called pseudoinverse or Moore-Penrose generalized inverse.

Substituting $w_{opt}$, we obtain the minimum cost function

$$J(w) = \frac{1}{2} |y - A w_{opt}|^2$$

$$= d^H d - d^H A (A^H A)^{-1} A^H d - d^H A (A^H A)^{-1} A^H d +$$

$$+ d^H A (A^H A)^{-1} (A^H d) (A^H A)^{-1} A^H d$$

$$= d^H d - d^H A (A^H A)^{-1} A^H d \quad (\star)$$

Define

$$E_d = d^H d = \sum_{i=1}^n |d_i|^2$$

$$E = E_d$$

$$E_{w_{opt}} = E_{w_{opt}} = \sum_{i=1}^n |e_i|^2$$

Consider

$$\hat{d} = A w_{opt} \quad \text{estimated output}$$

$$\hat{d}^H \hat{d} = w_{opt}^H A^H A w_{opt} =$$

$$= d^H A (A^H A)^{-1} A^H A (A^H A)^{-1} A^H d$$

$$= d^H A (A^H A)^{-1} A^H d \quad (\star \star) = \sum_{i=1}^n |\hat{d}_i|^2$$

$(\star \star)$ and $(\star)$ are the same.
Therefore at the point \( w = w_{opt} \)

\[ E_{win} = E_d - \hat{E}_d \]

\( E_{win} \) - energy of the error
\( E_d \) - energy of the desired input
\( \hat{E}_d \) - energy of the predicted output

\( \hat{E}_d / E_d \) - ratio of the desired output explained by the linear model ≤ 1

**Principle of Orthogonality Revised**

1° When \( w = w_{opt} \), the error vector is orthogonal to the columns of \( A \) matrix

\[ A^H e = A^H (d - \hat{d}) = A^H (d - A \cdot w_{opt}) = \]

\[ = A^H (d - A \cdot (A^H A)^{-1} A^H d) = A^H d - A^H A (A^H A)^{-1} A^H d \]

\[ = A^H d - A^H d = 0 \]

2° When \( w = w_{opt} \), \( \hat{d} \) and \( e \) are orthogonal.

\[ \hat{d}^H e = (A \cdot w_{opt})^H (d - A \cdot w_{opt}) = \]

\[ = w_{opt} A^H (d - A \cdot w_{opt}) = \]

\[ = d^H A (A^H A)^{-1} A^H (d - A \cdot (A^H A)^{-1} A^H d) = \]

\[ = d^H A (A^H A)^{-1} A^H d - d^H A (A^H A)^{-1} A^H A (A^H A)^{-1} A^H d \]

\[ = d^H A (A^H A)^{-1} A^H d - d^H A (A^H A)^{-1} A^H d = 0 \]
Physical interpretation

\[ \hat{d} = A \cdot w_{opt} = A \cdot (A^H A)^{-1} A^H \cdot d = P \cdot d \]

\( P \) - projector matrix that projects the desired vector onto the space spanned by columns of data matrix \( A \)

\[ P \cdot d \Rightarrow \text{part of the desired vector residing in the space spanned by columns of matrix } A \]

\[ I - P = I - (A (A^H A)^{-1} A^H) \Rightarrow \text{orthogonal complement projector} \]

Space orthogonal to the space spanned by \( A \) contains error vector

\[ d \rightarrow \hat{d} \]

\[ I - P \rightarrow e_{\text{min}} \]