

EE6617: 02 Linear Minimum Variance Unbiased (LMVU) Estimators

In this lecture note, we will study **linear MVU estimators**. This means we restrict the class of estimator to be not only unbiased but also linear.

Major reference:

Fundamentals of Statistical Processing, Volume I: Estimation Theory, 1/E, by Steven M. Kay, Prentice Hall, 1993

Statistical Signal Processing, 1/E, by Louis L. Scharf, Prentice Hall, 1991

1 Linear Models

While the CRLB is **very useful** (in that it provides a methodology to design MVU estimator), **determining MVU estimator is very difficult**.

Many **practical** problem can be well **approximated as linear** problems

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \quad (1)$$

$\mathbf{x} \in \mathbb{R}^{N \times 1}$: observation

$\mathbf{H} \in \mathbb{R}^{N \times p}$: known matrix

$\boldsymbol{\theta} \in \mathbb{R}^{p \times 1}$: unknown parameter vector

$\mathbf{w} \in \mathbb{R}^{N \times 1}$: noise vector, statistics known as $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$

Given the estimation parameter vector $\boldsymbol{\theta}$, the iff condition of CRLB says

$$\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta}) (\hat{\boldsymbol{\alpha}}(\mathbf{x}) - \boldsymbol{\theta}). \quad (2)$$

The likelihood function of $\boldsymbol{\theta}$ in (1) is

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2\right)$$

Let us first attempt to find the expression in (2).

$$\begin{aligned} \ln(p(\mathbf{x}; \boldsymbol{\theta})) &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 \\ \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} &= -\frac{1}{2\sigma^2} \frac{\partial \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2}{\partial \boldsymbol{\theta}} \\ \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) = \mathbf{x}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta} \end{aligned}$$

Note that

$$\begin{aligned}\frac{\partial \mathbf{x}^T \mathbf{H} \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} (\mathbf{x}^T \mathbf{H} \boldsymbol{\theta})^T = \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{x} = \mathbf{H}^T \mathbf{x} \\ \frac{\partial \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} &= 2 \frac{\partial}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta})^T = 2 \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta} = 2 \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \boldsymbol{\theta}} (-2\mathbf{x}^T \mathbf{H} \boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta})^T \\ &= \frac{1}{\sigma^2} (\mathbf{H}^T \mathbf{x} - \mathbf{H}^T \mathbf{H} \boldsymbol{\theta})\end{aligned}$$

If $\mathbf{H}^T \mathbf{H}$ is **invertible**, we can formulate in the similar form as (2),

$$\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{\mathbf{H}^T \mathbf{H}}{\sigma^2} \left((\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} - \boldsymbol{\theta} \right) = \mathbf{I}(\boldsymbol{\theta}) (\hat{\boldsymbol{\alpha}}(\mathbf{x}) - \boldsymbol{\theta})$$

In order to claim that $\hat{\boldsymbol{\alpha}}(\mathbf{x}) = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$ is the **efficient MVU estimator**, we need to check the **regularity condition** first

$$E \left[\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right] = \frac{1}{\sigma^2} (\mathbf{H}^T E[\mathbf{x}] - \mathbf{H}^T \mathbf{H} \boldsymbol{\theta}) = \mathbf{0}.$$

Thus, $\hat{\boldsymbol{\alpha}}(\mathbf{x}) = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$ is **efficient** and $\mathbf{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H}$ with the error covariance matrix

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = E \left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right)^T \right] = \mathbf{I}(\boldsymbol{\theta})^{-1} = \sigma^2 (\mathbf{H}^T \mathbf{H})^{-1}.$$

All of these results relies on $\mathbf{H}^T \mathbf{H}$ being **invertible** and this requires $N \geq p$. Thus, $\mathbf{H} \in \mathbb{R}^{N \times p}$ must be a **full rank square or tall matrix**. We summarize above result in the theorem.

Theorem 1 *Given the signal model in (1), the MVU estimator is*

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \mathbf{H}^\dagger \mathbf{x} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \\ \mathbf{C}_{\hat{\boldsymbol{\theta}}} &= \sigma^2 (\mathbf{H}^T \mathbf{H})^{-1} \\ \mathbf{I}(\boldsymbol{\theta}) &= \frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H}\end{aligned}$$

and it is *efficient*.

Example 1: Polynomial Fitting

Suppose that we have N measurements $x[t_n]$, $n = 0, \dots, N-1$ and the **structure of parameters $\boldsymbol{\theta}$ embedded in $\mathbf{x}$ is unknown**. Here t_n denotes the sampling time instance. The goal is to estimate $\hat{\boldsymbol{\theta}}$ to reconstruct $\mathbf{x}$. With the unknown structure of $\boldsymbol{\theta}$, we accommodate the linear model in (1). In particular, we model the measurement as $p-1$ th order polynomial such that

$$\begin{aligned}x[t_n] &= \theta_1 + \theta_2 t_n + \theta_3 t_n^2 + \dots + \theta_p t_n^{p-1} + w[t_n], \quad n = 0, \dots, N-1 \\ &= [1 \ t_n \ \dots \ t_n^{p-1}] \boldsymbol{\theta} + w[t_n].\end{aligned}$$

By collecting $x[t_0], \dots, x[t_{N-1}]$ in a vector,

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where

$$\mathbf{x} = \begin{bmatrix} x[t_0] \\ \vdots \\ x[t_{N-1}] \end{bmatrix} \quad \text{and} \quad \mathbf{H} = \begin{bmatrix} 1 & t_0 & \cdots & t_0^{p-1} \\ 1 & t_1 & \cdots & t_1^{p-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & t_{N-1} & \cdots & t_{N-1}^{p-1} \end{bmatrix} \in \mathbb{R}^{N \times p}, \quad N \geq p$$

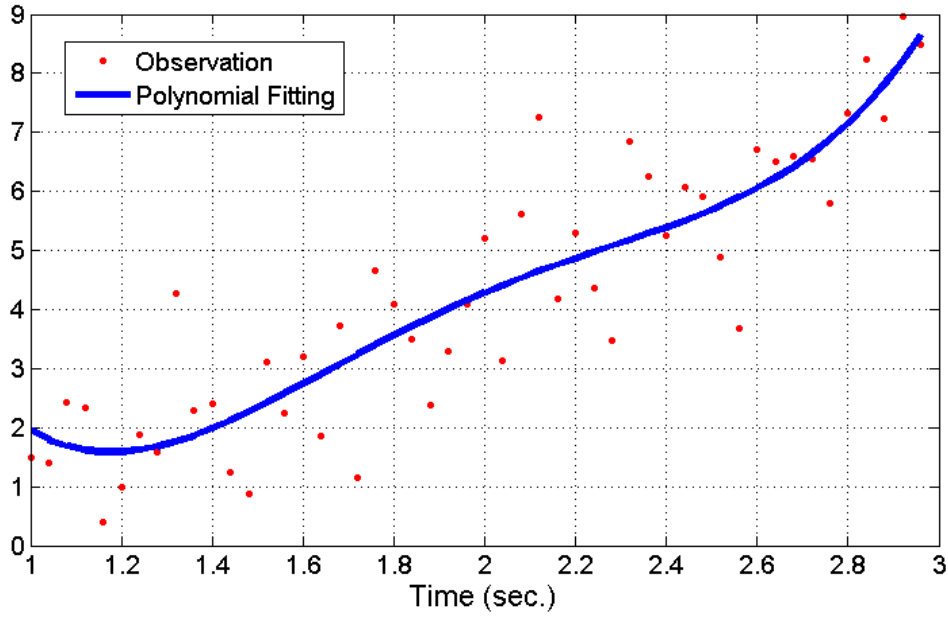
Then, the linear MVU estimator can be found as

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \quad (3)$$

and the reconstruction of $\mathbf{x}$, i.e., the polynomial fitting, can be given by

$$\mathbf{x}_{reconst} = \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}.$$

The linear filter $(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$ is often called **zero-forcing filter**.



Matlab Code Example:

```
close all;
clear all;
```

```
N = 50; % number of observations
p = 4; % order of polynomial
```

```

int = 0.04; % time sample interval
t = [0:int: int*(N-1)]+1; % time vector

% generate observation
w = randn(N,1); % generate zero mean and unit variance white Gaussian noise
x = t.^ 2 + w;

H = repmat(t', 1, p+1).^ repmat([0:p],N, 1); % H matrix
htheta = inv(H'*H)*H'*x; % estimate theta
hx = H*htheta; % reconstruct x, i.e., polynomial fitting

figure(1);
plot(t, x, 'r', t, hx, 'b'); grid on;

```

2 Linear Model with Correlated Noise

Now, we consider the scenario when the **noise is not i.i.d.** but **correlated, i.e., $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$** . Here we assume the noise covariance $E[\mathbf{w}\mathbf{w}^T] = \mathbf{C}$ is **positive definite $\mathbf{C} > 0$** . We want to estimate $\boldsymbol{\theta}$ from the linear model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}. \quad (4)$$

We will discuss two approaches. The first approach is based on the **estimation theory (CRLB)** and the second approach is based on the **communication signal processing (Noise Whitening)**.

2.1 CRLB Approach

Let us first apply CRLB. The likelihood function of $\boldsymbol{\theta}$ is

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2} \sqrt{\det(\mathbf{C})}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \right)$$

and attempt to find an expression in the form of $\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta}) (\hat{\boldsymbol{\alpha}}(\mathbf{x}) - \boldsymbol{\theta})$. To fulfill this, we compute

$$\begin{aligned}
\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} &= -\frac{1}{2} \frac{\partial (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= -\frac{1}{2} \frac{\partial (\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} - 2\mathbf{x}^T \mathbf{C}^{-1} \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} - \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\boldsymbol{\theta} \\
&= (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}) \left((\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} - \boldsymbol{\theta} \right).
\end{aligned}$$

Check the regularity condition $E \left[\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right] = \mathbf{H}^T \mathbf{C}^{-1} E[\mathbf{x}] - \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\boldsymbol{\theta} = \mathbf{0}$. Now, by using the **CRLB theorem**, we have the **Fisher information matrix**

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H},$$

the **linear MVU estimator**

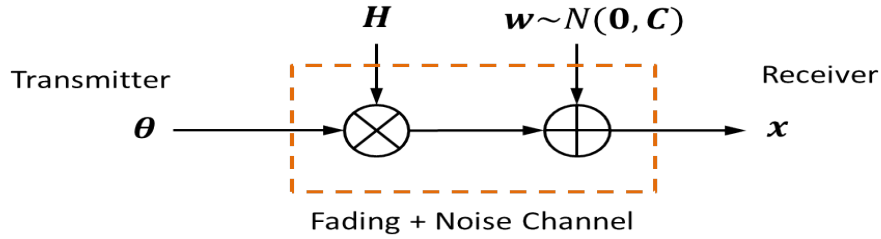
$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x},$$

and the **error covariance matrix**

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}.$$

2.2 Communication Signal Processing Approach

The signal model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ can be viewed as a **typical channel input and output signal model in Communications**.



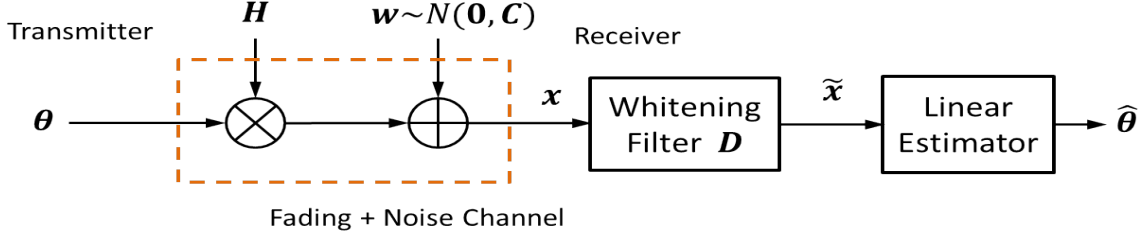
The matrix $\mathbf{H}$ is the **fading channel matrix**, $\boldsymbol{\theta}$ is the **transmit signal vector**, and $\mathbf{w}$ is the **additive colored noise**.

In order to correctly **estimate (or detect)** all the parameters in $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \cdots \ \theta_p]^T$ at the receiver, it is very crucial that the faded channel output $\mathbf{H}\boldsymbol{\theta}$ is received under an **evenly distributed noise**. Then, the receiver, given that the receiver has the **knowledge of channel $\mathbf{H}$** , can properly design a linear estimator as a function of $\mathbf{x}$ and $\mathbf{H}$, as given in Section 1.

On the other hand, if the noise is **colored (i.e., correlated)**, applying any genuine estimator would result in **noise amplification**. In this case, the **performance of the estimator (or detector) largely depends on a few of heavily contaminated signals** by noise.

Suppose that, for instance, the noise covariance matrix has one very large eigenvalue λ_1 and eigenvector $\mathbf{v}_1$ such that $\lambda_1 \gg \lambda_2 \approx \lambda_3 \approx \cdots \approx \lambda_N \approx 0$. If a parameter θ_1 experiences the channel fading in the direction of $\mathbf{v}_1$, i.e., if the first column of $\mathbf{H}$ is $\mathbf{v}_1 \in \mathbb{R}^{N \times 1}$, estimating (or detecting) the parameter θ_1 is very challenging compared to the other parameters, and it creates large errors. Furthermore, it requires complicated estimators, e.g., non-linear estimator, iterative algorithms, etc, in order to combat with the biased noise power.

In the communication signal processing, this difficulty can be overcome by adopting the **linear noise whitening filter**.



If the receiver has **knowledge about the noise covariance matrix $\mathbf{C}$** , the receiver could design the linear filter $\mathbf{D}$ to satisfy some desired property. Given the whitening filter, write

$$\begin{aligned}\tilde{\mathbf{x}} &= \mathbf{D}\mathbf{x} \\ &= \mathbf{D}\mathbf{H}\boldsymbol{\theta} + \mathbf{D}\mathbf{w} \\ &= \tilde{\mathbf{H}}\boldsymbol{\theta} + \tilde{\mathbf{w}}.\end{aligned}\tag{5}$$

The **goal is to design $\mathbf{D}$ to whiten the noise $\mathbf{w}$ such that**

$$E[\tilde{\mathbf{w}}\tilde{\mathbf{w}}^T] = E[\mathbf{D}\mathbf{w}\mathbf{w}^T\mathbf{D}^T] = \mathbf{D}E[\mathbf{w}\mathbf{w}^T]\mathbf{D}^T = \mathbf{D}\mathbf{C}\mathbf{D}^T = \mathbf{I}.$$

Suppose $\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D}$. Then, $\mathbf{D}\mathbf{C}\mathbf{D}^T = \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T = \mathbf{D}\mathbf{D}^{-1}\mathbf{D}^{-T}\mathbf{D}^T = \mathbf{I}$, implying the **whitening filter can be designed by decomposing $\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D}$** .

How can we attain the decomposition $\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D}$? Since $\mathbf{C} > 0$ and $\mathbf{C}$ is symmetric, $\mathbf{C}^{-1} > 0$ and $\mathbf{C}^{-1}$ is also symmetric. Thus, all the eigenvalues (or singular values) of $\mathbf{C}^{-1}$ are positive and the EVD (or SVD) of $\mathbf{C}^{-1}$ can be written as

$$\mathbf{C}^{-1} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T = \underbrace{\mathbf{V}\boldsymbol{\Lambda}^{1/2}}_{\mathbf{D}^T} \underbrace{\boldsymbol{\Lambda}^{1/2}\mathbf{V}^T}_{\mathbf{D}} = \mathbf{D}^T\mathbf{D}$$

and the **whitening filter is given by $\mathbf{D} = \boldsymbol{\Lambda}^{1/2}\mathbf{V}^T$** .

From the signal model in (5), now we can apply Theorem 1 to design efficient linear MVU estimator

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\tilde{\mathbf{H}}^T\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}^T\tilde{\mathbf{x}} \\ &= (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} \\ \mathbf{I}(\hat{\boldsymbol{\theta}}) &= \tilde{\mathbf{H}}^T\tilde{\mathbf{H}} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H} \\ \mathbf{C}_{\hat{\boldsymbol{\theta}}} &= (\tilde{\mathbf{H}}^T\tilde{\mathbf{H}})^{-1} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}.\end{aligned}$$

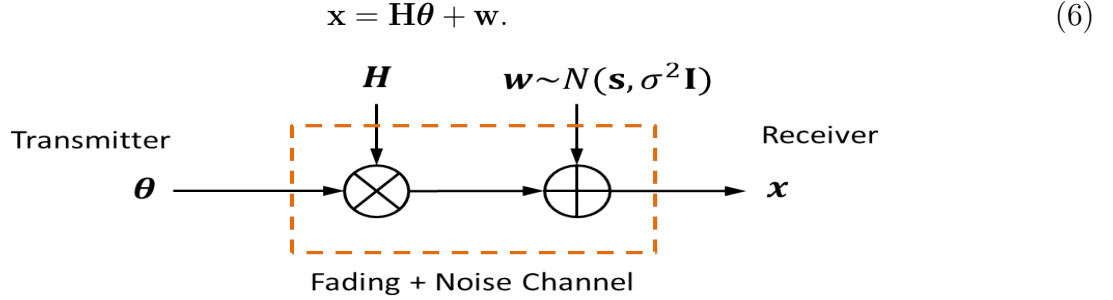
You can check this whitening approach yields the same MVU estimator as the CRLB approach in Section 2.1.

Again, the **main idea is to manipulate the received signal so that the received noise $\mathbf{w}$ has the same statistics as $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ in (1)**. Then, apply the Theorem 1 to obtain the efficient MVU estimator.

As you may see, this whitening approach provides more **intuitive strategy** and in many of cases, it is easier than CRLB approach for designing the **linear efficient estimator**.

3 Linear Model with Non-zero Mean Noise

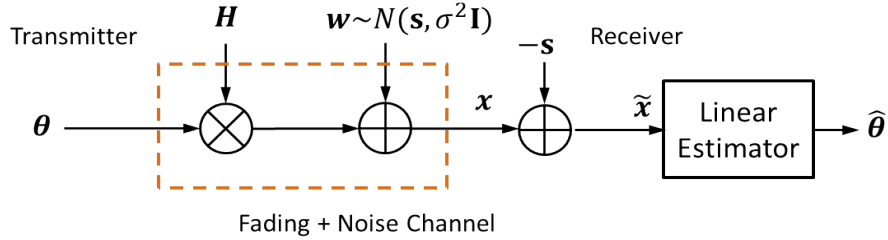
In this section, we consider the case when the **noise is biased but not correlated, i.e., $\mathbf{w} \sim \mathcal{N}(\mathbf{s}, \sigma^2 \mathbf{I})$, $\mathbf{s} \neq \mathbf{0}$.**



Similar to Section 2.2, we could equalize the biased noise to be zero mean by adopting **linear processing** at the receiver. This can be done by

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{x} - \mathbf{s} \\ &= \mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{s} \\ &= \mathbf{H}\boldsymbol{\theta} + \tilde{\mathbf{w}} \end{aligned} \quad (7)$$

where $\tilde{\mathbf{w}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.



Since (7) has the same statistics as (1), the **efficient MVU estimator** is found as

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \tilde{\mathbf{x}} \\ &= (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T (\mathbf{x} - \mathbf{s}) \end{aligned}$$

and the error covariance matrix is given by

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2 (\mathbf{H}^T \mathbf{H})^{-1}.$$

Can you design the efficient linear MVU estimator given the observation $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{s}, \mathbf{C})$?