Minimum Mean-Square Error (MMSE) and Linear MMSE (LMMSE) Estimation

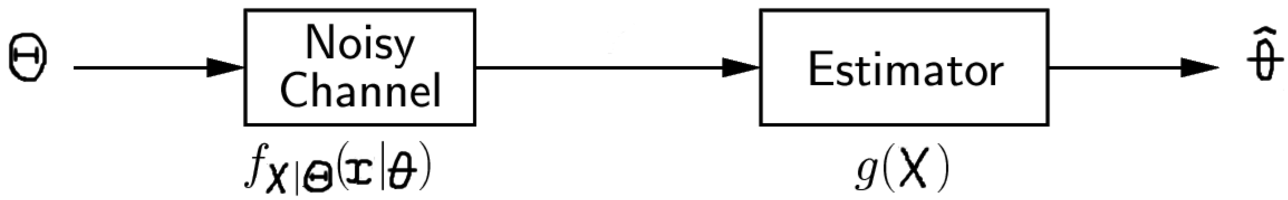
Outline:

- MMSE estimation,
- Linear MMSE (LMMSE) estimation,
- Geometric formulation of LMMSE estimation and orthogonality principle.

Reading: Chapter 12 in Kay-I.

MMSE Estimation

Consider the following problem:



A signal $\Theta = \theta$ is transmitted through a noisy channel, modeled using the conditional pdf $f_{X\,|\,\Theta}(x\,|\,\theta)$, which is the likelihood function of θ. We observe $X = x$. The signal Θ has known *prior (marginal) pdf*

$$f_\Theta(\theta) = \pi(\theta)$$

which summarizes our knowledge about Θ before (i.e. *prior to*) collecting $X = x$. We wish to *estimate* Θ using the observation $X = x$:

$$\widehat{\theta} = \widehat{\theta}(x) = g(x).$$

We choose $g(X)$ to minimize the *"Bayesian" (preposterior) mean-square error*:

$$\text{BMSE} = \text{E}_{\,\Theta,X}\{[\widehat{\theta}(X) - \Theta]^2\} = \text{E}_{\,\Theta,X}\{[g(X) - \Theta]^2\}.$$

Here, $\widehat{\theta}(X)$ that achieve the minimum BMSE are called *minimum MSE (MMSE) estimates* of Θ. $\widehat{\theta}(X)$ may not be unique.

A Reminder: MMSE Estimation

Theorem 1. *The MMSE estimate of Θ (based on the observation $X = x$) is given by*

$$\widehat{\theta}_{\rm MMSE} = g(x) = {\rm E}_{\,\Theta \,|\, X}(\Theta \,|\, x). \tag{1}$$

The minimum BMSE (i.e. the BMSE of $\widehat{\theta}_{\rm MMSE}(x) = {\rm E}_{\,\Theta \,|\, X}[\Theta \,|\, x]$) is

$$\begin{aligned} \text{MBMSE} &= {\rm E}_{\,X}[\text{var}_{\Theta \,|\, X}(\Theta \,|\, X)] \\ &= {\rm E}_{\,\Theta}(\Theta^2) - {\rm E}_{\,X}\{[{\rm E}_{\,\Theta \,|\, X}(\Theta \,|\, X)]^2\}. \end{aligned} \tag{2}$$

Lemma 1. *We first show that*

$$\min_b {\rm E}_{\,\Theta}[(b - \Theta)^2] = \text{var}_{\Theta}(\Theta)$$

is achieved for

$$b = {\rm E}_{\,\Theta}(\Theta).$$

Therefore, in absence of any observations, the MMSE estimate *of Θ is equal to the mean of the (prior, marginal) pdf of Θ:*

$$\begin{aligned}
\mathrm{E}_{\Theta}[(\Theta - b)^2] &= \mathrm{E}_{\Theta}[(\Theta - \mathrm{E}_{\Theta}(\Theta) + \mathrm{E}_{\Theta}(\Theta) - b)^2] \\
&= \mathrm{E}_{\Theta}\Big\{[\Theta - \mathrm{E}_{\Theta}(\Theta)]^2 + [\mathrm{E}_{\Theta}(\Theta) - b]^2 \\
&\quad +2\,[\mathrm{E}_{\Theta}(\Theta) - b]\,[\Theta - \mathrm{E}_{\Theta}(\Theta)]\Big\} \\
&= \mathrm{E}_{\Theta}[(\Theta - \mathrm{E}_{\Theta}[\Theta])^2] + (\mathrm{E}_{\Theta}[\Theta] - b)^2 \\
&\quad +2\,[\mathrm{E}_{\Theta}(\Theta) - b]\,\underbrace{\mathrm{E}_{\Theta}[\Theta - \mathrm{E}_{\Theta}(\Theta)]}_{0} \\
&\geq \mathrm{E}_{\Theta}\{[\Theta - \mathrm{E}_{\Theta}(\Theta)]^2\}
\end{aligned}$$

with equality if and only if $b = \mathrm{E}_{\Theta}(\Theta)$.

Proof. **(Theorem 1)** We now consider our MMSE estimation problem, write BMSE of an estimator $g(X)$ as

$$\begin{aligned}
\mathrm{BMSE} &= \mathrm{E}_{\Theta,X}\{[\Theta - g(X)]^2\} \\
&\overset{\text{iter. exp.}}{=} \mathrm{E}_{X}\Big(\underbrace{\mathrm{E}_{\Theta\,|\,X}\{[\Theta - g(X)]^2\,|\,X\}}_{\rho(\widehat{\theta}\,|\,X),\ \text{see handout \# 4}}\Big)
\end{aligned}$$

and use Lemma 1 to conclude that, for each $X = x$, the posterior expected squared loss

$$\rho(\widehat{\theta}\,|\,x) = \mathrm{E}_{\Theta\,|\,X}\{[\Theta - g(X)]^2\,|\,x\}$$

is minimized for

$$g(x) = \mathrm{E}_{\Theta \,|\, X}(\Theta \,|\, x).$$

Thus, BMSE is minimized for

$$g(X) = \mathrm{E}_{\Theta \,|\, X}(\Theta \,|\, X).$$

We now find the minimum BMSE:

$$\begin{aligned} \mathrm{MBMSE} &= \mathrm{E}_{\Theta, X}\{[\Theta - \mathrm{E}_{\Theta \,|\, X}(\Theta \,|\, X)]^2\} \\ &\overset{\text{iter. exp.}}{=} \mathrm{E}_{X}[\mathrm{E}_{\Theta \,|\, X}\{[\Theta - \mathrm{E}_{\Theta \,|\, X}(\Theta \,|\, X)]^2 \,|\, X\}] \\ &= \mathrm{E}_{X}[\mathrm{var}_{\Theta \,|\, X}(\Theta \,|\, X)]. \end{aligned} \tag{3}$$

□

Comments:

-

$$\mathrm{E}_{X}[\widehat{\theta}_{\mathrm{MMSE}}(X)] = \mathrm{E}_{\Theta}(\Theta) \quad \text{unbiased on average.} \tag{4}$$

However, $\widehat{\theta}_{\mathrm{MMSE}}(X)$ is practically never unbiased in the classical sense:

$$\mathrm{E}_{X \,|\, \Theta}[\widehat{\theta}_{\mathrm{MMSE}}(X) \,|\, \theta] \neq \theta \quad \text{in general.} \tag{5}$$

You will show (5) in a HW assignment.

- For *independent* Θ and X, the MMSE estimate of Θ is

$$\widehat{\theta}_{\rm MMSE}(X) = E_\Theta(\Theta).$$

- The estimation error

$$\mathcal{E} = \widehat{\theta}_{\rm MMSE}(X) - \Theta \tag{6}$$

and the MMSE estimate $\widehat{\theta}_{\rm MMSE}(X)$ are *orthogonal*:

$$\begin{aligned}
\mathrm{E}_{\,\Theta,X}[\mathcal{E}\,\widehat{\theta}_{\rm MMSE}(X)] &= \mathrm{E}_{\,\Theta,X}\{[\widehat{\theta}_{\rm MMSE}(X) - \Theta]\,\widehat{\theta}_{\rm MMSE}(X)\} \\
&\overset{\text{iter. exp.}}{=} \mathrm{E}_{\,X}\{\mathrm{E}_{\,\Theta\,|\,X}([\widehat{\theta}_{\rm MMSE}(X) - \Theta]\,\widehat{\theta}_{\rm MMSE}(X)\,|\,X)\} \\
&= \mathrm{E}_{\,X}\{\widehat{\theta}_{\rm MMSE}(X)\,\mathrm{E}_{\,\Theta\,|\,X}[\Theta - \widehat{\theta}_{\rm MMSE}(X)\,|\,X]\} = 0
\end{aligned}$$

since $\widehat{\theta}_{\rm MMSE}(X) = \mathrm{E}_{\,\Theta\,|\,X}[\Theta\,|\,X]$. It is clear from this derivation that the estimation error $\mathcal{E}$ in (6) is orthogonal to *any* function $g(X)$ of X:

$$\begin{aligned}
&\mathrm{E}_{\,\Theta,X}\{[\Theta - \widehat{\theta}_{\rm MMSE}(X)]\,g(X)\} \\
&= \mathrm{E}_{\,X}\{\mathrm{E}_{\,\Theta\,|\,X}([\Theta - \widehat{\theta}_{\rm MMSE}(X)]\,g(X)\,|\,X)\} \\
&= \mathrm{E}_{\,X}\{g(X)\,\mathrm{E}_{\,\Theta\,|\,X}[\Theta - \widehat{\theta}_{\rm MMSE}(X)\,|\,X]\} = 0.
\end{aligned}$$

- The law of conditional variances [(5) in handout # 0b] implies

$$\mathrm{var}_{\Theta}(\Theta) = \underbrace{\mathrm{E}_{X}[\mathrm{var}_{\Theta \mid X}(\Theta \mid X)]}_{\text{MBMSE, see (3)}} + \mathrm{var}_{X}(\underbrace{\mathrm{E}_{\Theta \mid X}[\Theta \mid X]}_{\widehat{\theta}_{\mathrm{MMSE}}(X), \text{ see (1)}})$$

i.e. the sum of

- the minimum BMSE for estimating Θ and
- variance of the MMSE estimate of Θ

is equal to the (marginal, prior) variance of Θ.

Additive Gaussian Noise Channel

Consider a communication channel with input

$$\Theta \sim \mathcal{N}(\mu_\Theta, \tau_\Theta^2)$$

and noise

$$W \sim \mathcal{N}(0, \sigma^2)$$

where Θ and W are independent and the measurement X is modeled as

$$X = \Theta + W. \tag{7}$$

Find the MMSE estimate of Θ based on X and the resulting minimum BMSE (MBMSE), i.e. $\mathrm{E}_{\Theta \mid X}(\Theta \mid X)$ and $\mathrm{E}_X[\mathrm{var}_{\Theta \mid X}(\Theta \mid X)]$, see (1) and (2).

Note: We have already considered this problem in handout # 4. We revisit it here with focus on MMSE estimation and finding MBMSE.

Solution: From (7), we have:

$$f_{X \mid \Theta}(x \mid \theta) = \mathcal{N}(x \mid \theta, \sigma^2).$$

We now find $f_{\Theta \mid X}(\theta \mid x)$ using the Bayes' rule:

$$
\begin{aligned}
f_{\Theta \mid X}(\theta \mid x) &\propto f_{\Theta}(\theta)\, f_{X \mid \Theta}(x \mid \theta) \\
&\propto \exp\Big[-\frac{1}{2\,\tau_\Theta^2}\,(\theta-\mu_\Theta)^2\Big]\cdot \exp\Big[-\frac{1}{2\,\sigma^2}\,(x-\theta)^2\Big] \\
&\propto \exp\Big[-\tfrac{1}{2}\,(\frac{1}{\tau_\Theta^2}+\frac{1}{\sigma^2})\,\theta^2 + (\frac{1}{\tau_\Theta^2}\,\mu_\Theta + \frac{1}{\sigma^2}\,x)\,\theta\Big] \\
&= \mathcal{N}\left(\theta \,\Big|\, \frac{\frac{1}{\sigma^2}\,x + \frac{1}{\tau_\Theta^2}\,\mu_\Theta}{\frac{1}{\sigma^2}+\frac{1}{\tau_\Theta^2}}, \Big(\frac{1}{\sigma^2}+\frac{1}{\tau_\Theta^2}\Big)^{-1}\right)
\end{aligned}
$$

implying that

$$
\widehat{\theta}_{\mathrm{MMSE}}(X) = \mathrm{E}_{\,\Theta \mid X}(\Theta \mid X) = \frac{\frac{1}{\sigma^2}\,X + \frac{1}{\tau_\Theta^2}\,\mu_\Theta}{\frac{1}{\sigma^2}+\frac{1}{\tau_\Theta^2}} \qquad (8)
$$

$$
\mathrm{var}_{\Theta \mid X}(\Theta \mid X) = \Big(\frac{1}{\sigma^2}+\frac{1}{\tau_\Theta^2}\Big)^{-1} \qquad (9)
$$

and, consequently,

$$
\mathrm{MBMSE} = \mathrm{E}_{\,X}[\mathrm{var}_{\Theta \mid X}(\Theta \mid X)] = \Big(\frac{1}{\sigma^2}+\frac{1}{\tau_\Theta^2}\Big)^{-1}. \qquad (10)
$$

Note: In the above example, the MMSE estimate is a linear (more precisely, constant + linear = affine) function of the observation X. This is not always the case, e.g. for

$$f_{\Theta \mid X}(\theta \mid x) = \begin{cases} x\, e^{-x\,\theta} & x > 0, \theta \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

we obtain $\mathrm{E}_{\Theta \mid X}(\Theta \mid X) = 1/X$. Here is another example.

Computing the MMSE estimator: Another example.

Example: Let

$$f_{\Theta,X}(\theta,x) = \begin{cases} 2, & \text{if } \theta \geq 0, x \geq 0, \theta + x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

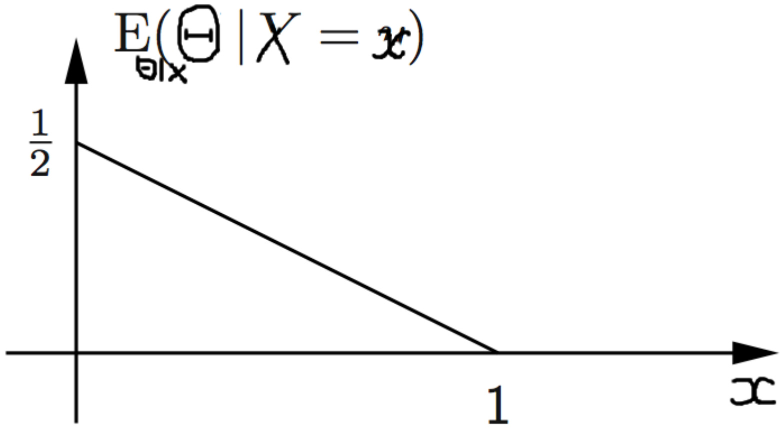
Find $\mathrm{E}_{\Theta|X}(\Theta \mid X = x)$.

Solution: We already know that

$$f_{\Theta|X}(\theta|x) = \begin{cases} \frac{1}{1-x} & \text{if } \theta \geq 0, x \geq 0, \theta + x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

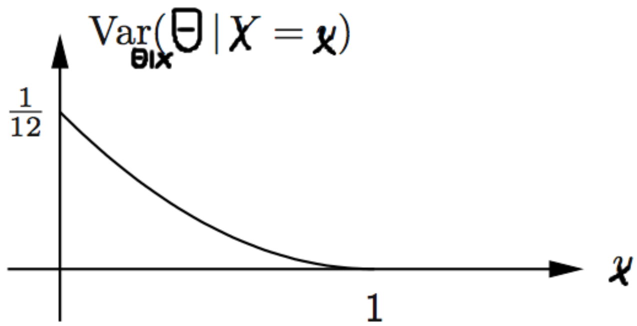
Thus

$$\begin{aligned} \mathrm{E}_{\Theta|X}(\Theta|X = x) &= \int_0^{1-x} \frac{1}{1-x} \theta \, d\theta \\ &= \frac{1-x}{2}, \quad 0 \leq x < 1. \end{aligned}$$



And, for $X = x$, the minimum MSE is given by

$$\mathrm{Var}_{\Theta|X}(\Theta|X = x) = \frac{(1-x)^2}{12},\ 0 \le x < 1$$



Thus the minimum MSE is $E_X(\mathrm{Var}_{\Theta|X}(\Theta\,|\,X)) = \frac{1}{24}$, compared to $\mathrm{Var}_{\Theta}(\Theta) = \frac{1}{18}$

The difference is $\mathrm{Var}_X(E_{\Theta|X}(\Theta|X)) = \frac{1}{72}$, which is the variance of the estimate.

Gaussian Linear Model (Theorem 10.3 in Kay-I)

Theorem 2. *Consider the linear model:*

$$\boldsymbol{X} = H\,\boldsymbol{\theta} + \boldsymbol{W}$$

H *is a known matrix, and*

$$\begin{aligned} \boldsymbol{W} &\sim \mathcal{N}(\mathbf{0}, C_{\boldsymbol{W}}) \\ \boldsymbol{\Theta} &\sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\Theta}}, C_{\boldsymbol{\Theta}}) \end{aligned}$$

where $\boldsymbol{W}$ *and* $\boldsymbol{\Theta}$ *are independent and* $C_{\boldsymbol{W}}, \boldsymbol{\mu}_{\boldsymbol{\Theta}}$, *and* $\boldsymbol{C}_{\boldsymbol{\Theta}}$ *are known hyperparameters. Then, the posterior pdf* $f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ *is Gaussian:*

$$f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \mathcal{N}\Big(\boldsymbol{\theta}\,|\,(H^T C_{\boldsymbol{W}}^{-1} H + C_{\boldsymbol{\Theta}}^{-1})^{-1}\,(H^T C_{\boldsymbol{W}}^{-1}\boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\Theta}}), \\ (H^T\,C_{\boldsymbol{W}}^{-1}\,H + C_{\boldsymbol{\Theta}}^{-1})^{-1}\Big). \tag{11}$$

Proof.

$$
\begin{aligned}
f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) &\propto f_{\boldsymbol{X}\,|\,\boldsymbol{\Theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}) \\
&\propto \exp[-\tfrac{1}{2}\,(\boldsymbol{x} - H\,\boldsymbol{\theta})^T\,C_{\boldsymbol{W}}^{-1}\,(\boldsymbol{x} - H\,\boldsymbol{\theta})] \\
&\quad \cdot \exp[-\tfrac{1}{2}\,(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\Theta}})^T\,C_{\boldsymbol{\Theta}}^{-1}\,(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\Theta}})] \\
&\propto \exp(-\tfrac{1}{2}\,\boldsymbol{\theta}^T\,H^T\,C_{\boldsymbol{W}}^{-1}\,H\boldsymbol{\theta} + \boldsymbol{x}^T\,C_{\boldsymbol{W}}^{-1}\,H\,\boldsymbol{\theta}) \\
&\quad \cdot \exp(-\tfrac{1}{2}\,\boldsymbol{\theta}^T\,C_{\boldsymbol{\Theta}}^{-1}\,\boldsymbol{\theta} + \boldsymbol{\mu}_{\boldsymbol{\Theta}}^T\,C_{\boldsymbol{\Theta}}^{-1}\,\boldsymbol{\theta}) \\
&= \exp[-\tfrac{1}{2}\,\boldsymbol{\theta}^T\,(H^T\,C_{\boldsymbol{W}}^{-1}\,H + C_{\boldsymbol{\Theta}}^{-1})\,\boldsymbol{\theta} + (\boldsymbol{x}^T\,C_{\boldsymbol{W}}^{-1}\,H + \boldsymbol{\mu}_{\boldsymbol{\Theta}}^T\,C_{\boldsymbol{\Theta}}^{-1})\,\boldsymbol{\theta}] \\
&\propto \mathcal{N}\Big((H^T\,C_{\boldsymbol{W}}^{-1}\,H + C_{\boldsymbol{\Theta}}^{-1})^{-1}\,(\boldsymbol{H}^T C_{\boldsymbol{W}}^{-1}\boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\Theta}}), \\
&\quad (H^T\,C_{\boldsymbol{W}}^{-1}\,H + C_{\boldsymbol{\Theta}}^{-1})^{-1}\Big).
\end{aligned}
$$

□

Comments:

- DC-level estimation in AWGN with known variance introduced on p. 17 of handout # 4 is a special case of this result, see also Example 10.2 in Kay-I.

- Examine the posterior mean:

$$\mathrm{E}_{\Theta\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = (\underbrace{H^T C_{\boldsymbol{W}}^{-1} H}_{\text{likelihood precision}} + \underbrace{C_{\boldsymbol{\Theta}}^{-1}}_{\text{prior precision}})^{-1} \cdot (\underbrace{H^T C_{\boldsymbol{W}}^{-1} \boldsymbol{x}}_{\text{data-dependent term}} + \underbrace{C_{\boldsymbol{\Theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}}}_{\text{prior-dependent term}}).$$

- **Noninformative (flat) prior on θ and white noise.** Consider the Jeffreys' noninformative (flat) prior pdf for $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}) \propto 1 \quad (\boldsymbol{C}_{\boldsymbol{\theta}}^{-1} = \boldsymbol{0})$$

and white noise:

$$C_{\boldsymbol{W}} = \sigma^2 \underbrace{I}_{\text{identity matrix}} .$$

Then, $f_{\Theta\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ in (11) simplifies to

$$f_{\Theta\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \mathcal{N}\Big(\boldsymbol{\theta}\,\Big|\, \overbrace{(H^T H)^{-1} H^T \boldsymbol{x}}^{\widehat{\boldsymbol{\theta}}_{\mathrm{LS}(\boldsymbol{x})}},\ \sigma^2 (H^T H)^{-1}\Big).$$

- **Prediction:** We now practice prediction for this model. Say we wish to predict a $X_\star$ coming from the following model:

$$X_\star = \boldsymbol{h}_\star^T \boldsymbol{\theta} + W_\star$$

where $W_\star \sim \mathcal{N}(0, \sigma^2)$ is independent from $\boldsymbol{W}$, implying that $X_\star$ and $\boldsymbol{X}$ are conditionally independent given $\boldsymbol{\Theta} = \boldsymbol{\theta}$ and, therefore,

$$f_{X_\star \,|\, \boldsymbol{\Theta}, \boldsymbol{X}}(x_\star \,|\, \boldsymbol{\theta}, \boldsymbol{x}) = f_{X_\star \,|\, \boldsymbol{\Theta}}(x_\star \,|\, \boldsymbol{\theta}) = \mathcal{N}(x_\star \,|\, \boldsymbol{h}_\star^T \, \boldsymbol{\theta}, \sigma^2).$$

Then, our posterior predictive pdf is [along the lines of (10)]

$$f_{X_\star \,|\, \boldsymbol{X}}(x_\star \,|\, \boldsymbol{x}) = \int \underbrace{f_{X_\star \,|\, \boldsymbol{\Theta}}(x_\star \,|\, \boldsymbol{\theta})}_{\mathcal{N}(x_\star \,|\, \boldsymbol{h}_\star^T \boldsymbol{\theta}, \sigma^2)} \cdot \underbrace{f_{\boldsymbol{\Theta} \,|\, \boldsymbol{X}}(\boldsymbol{\theta} \,|\, \boldsymbol{x})}_{\mathcal{N}(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}), C_{\text{post}})} \; d\boldsymbol{\theta}$$

where

$$\begin{aligned} \widehat{\boldsymbol{\theta}}(\boldsymbol{x}) &= (H^T \, C_{\boldsymbol{W}}^{-1} \, H + C_{\boldsymbol{\Theta}}^{-1})^{-1} \, (H^T \, C_{\boldsymbol{w}}^{-1} \, \boldsymbol{x} + C_{\boldsymbol{\Theta}}^{-1} \, \boldsymbol{\mu}_{\boldsymbol{\Theta}}) \\ C_{\text{post}} &= (H^T \, C_{\boldsymbol{W}}^{-1} \, H + C_{\boldsymbol{\Theta}}^{-1})^{-1} \end{aligned}$$

implying

$$f_{X_\star \,|\, \boldsymbol{X}}(x_\star \,|\, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{h}_\star^T \, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{h}_\star^T \, C_{\text{post}} \, \boldsymbol{h}_\star + \sigma^2).$$

Linear MMSE (LMMSE) Estimation

For exact MMSE estimation, we need to know the joint pdf (or joint pmf) $f_{\Theta,X}(\theta, x)$, typically specified through the *prior (marginal) pdf/pmf* $f_\Theta(\theta)$ and *conditional pdf/pmf* $f_{X \,|\, \Theta}(x \,|\, \theta)$, which together yield the joint pdf (or pmf, or combined pdf/pmf)

$$f_{\Theta,X}(\theta, x) = f_{X \,|\, \Theta}(x \,|\, \theta)\, f_\Theta(\theta).$$

This information may not be available.

- We typically have estimates of the first and second moments of the signal and the observation, i.e. of the means, variances, and covariance between Θ and X.

- This information is generally *not sufficient* for MMSE estimation of Θ, but is sufficient for *linear MMSE (LMMSE) estimation* of Θ, i.e. for finding estimates of the form:

$$\widehat{\theta} = \widehat{\theta}(X) = a\, X + b \tag{12}$$

  that minimize *BMSE*:

$$\text{BMSE} = \text{E}_{\,\Theta,X}\{[\Theta - \widehat{\theta}(X)]^2\}.$$

  The minimization is with respect to a and b.

Note: Even though it is more appropriate to refer to this estimator as 'affine MMSE estimator,' linear MMSE estimator is the most widely used name for it. In most applications, we consider zero-mean X and Θ; then, our estimator is indeed linear, see Theorem 3 below.

Theorem 3. *The LMMSE estimate of Θ is*

$$
\begin{aligned}
\widehat{\theta}(X) &= \underbrace{\frac{\text{cov}_{\Theta,X}(\Theta, X)}{\sigma_X^2}}_{a_{\text{opt}}} \cdot [X - \mathrm{E}_X(X)] + \mathrm{E}_\Theta(\Theta) \\
&= \rho_{\Theta,X}\, \sigma_\Theta \frac{X - \mathrm{E}_X(X)}{\sigma_X} + \mathrm{E}_\Theta(\Theta) \qquad (13)
\end{aligned}
$$

and its BMSE is given by

$$
\begin{aligned}
\text{MBMSE}_{\text{linear}} &= \text{cov}_{\Theta,X}(\Theta - \widehat{\theta}(X), \Theta) \qquad (14) \\
&= \sigma_\Theta^2 - \frac{\text{cov}^2_{\Theta,X}(\Theta, X)}{\sigma_X^2} = (1 - \rho^2_{\Theta,X})\, \sigma_\Theta^2. \qquad (15)
\end{aligned}
$$

Here,

$$
\begin{aligned}
\text{cov}_{\Theta,X}(\Theta, X) &= \mathrm{E}_{\Theta,X}[(\Theta - \mu_\Theta)\,(X - \mu_X)] \\
&= \mathrm{E}_{\Theta,X}(\Theta\, X) - \mathrm{E}_\Theta(\Theta)\, \mathrm{E}_X(X)
\end{aligned}
$$

is introduced on p. 4 of handout # 0b,

$$\begin{aligned}\operatorname{var}_\Theta(\Theta) &= \operatorname{cov}_\Theta(\Theta,\Theta) = \sigma_\Theta^2\\ \operatorname{var}_X(X) &= \sigma_X^2\end{aligned}$$

and $\rho_{\Theta,X}$ is the correlation coefficient between Θ and X, defined as

$$\rho_{\Theta,X} = \frac{\operatorname{cov}_{\Theta,X}(\Theta, X)}{\sqrt{\operatorname{var}_\Theta(\Theta)}\cdot\sqrt{\operatorname{var}_X(X)}} = \frac{\operatorname{cov}_{\Theta,X}(\Theta, X)}{\sigma_\Theta\,\sigma_X}$$

where $\sigma_\Theta = \sqrt{\sigma_\Theta^2}$ and $\sigma_X = \sqrt{\sigma_X^2}$ are the (marginal) standard deviations of Θ and X.

Proof. Suppose first that the constant a has already been chosen. Then, choosing the constant b to minimize the BMSE

$$\mathrm{E}_{\Theta,X}[(\Theta - a\,X - b)^2]$$

is equivalent to finding b that minimizes $\mathrm{E}_\Xi[(\Xi - b)^2]$, where $\Xi \stackrel{\triangle}{=} \Theta - a\,X$. This problem is solved in Lemma 1, and the optimal b is $b = \mathrm{E}_\Xi(\Xi)$, i.e.

$$b = \mathrm{E}_\Xi(\Xi) = \mathrm{E}_{\Theta,X}(\Theta - a\,X) = \mathrm{E}_\Theta(\Theta) - a\,\mathrm{E}_X(X). \quad (16)$$

Substituting (16) into $\mathrm{E}_{\Theta,X}[(\underbrace{\Theta - a\,X}_{\Xi} - b)^2]$ yields:

$$\mathrm{E}_{\Xi}\{[\Xi - \mathrm{E}_{\Xi}(\Xi)]^2\} = \mathrm{var}_{\Xi}(\Xi) = \mathrm{var}_{\Theta,X}(\Theta - a\,X) \quad (17)$$
$$= \sigma_\Theta^2 + a^2\,\sigma_X^2 - 2\,a\,\mathrm{cov}_{\Theta,X}(\Theta, X) \quad (18)$$

which is easy to minimize with respect to a. In particular, differentiating (18) with respect to a and setting the result to zero yields

$$2\,a\,\sigma_X^2 - 2\,\mathrm{cov}_{\Theta,X}(\Theta, X) = 0$$

i.e.

$$a\,\mathrm{cov}_{\Theta,X}(X, X) - \mathrm{cov}_{\Theta,X}(\Theta, X) = 0$$

and, finally,

$$\mathrm{cov}_{\Theta,X}(a\,X - \Theta,\ X) = 0$$

which is the famous *orthogonality principle*. Clearly, the optimal a is

$$a_{\mathrm{opt}} = \frac{\mathrm{cov}_{\Theta,X}(\Theta, X)}{\sigma_X^2} \quad (19)$$

and (13) follows. We summarize the orthogonality principle:

$$\text{cov}_{\Theta,X}(a_{\text{opt}}\,X - \Theta, X) = 0 \tag{20}$$

or, equivalently,

$$\text{cov}_{\Theta,X}\Big(\underbrace{\widehat{\theta}(X)}_{\substack{\text{LMMSE est. of } \Theta \\ \text{based on } X}} - \Theta, X\Big) = 0. \tag{21}$$

Substituting (19) into (17) yields

$$\begin{aligned}
\text{MMSE}_{\text{linear}} &= \underbrace{\text{cov}_{\Theta,X}(\Theta - a_{\text{opt}}\,X, \Theta - a_{\text{opt}}\,X)}_{\text{var}_{\Theta,X}(\Theta - a_{\text{opt}}\,X)} \\
&= \text{cov}_{\Theta,X}(\Theta - a_{\text{opt}}\,X, \Theta) - a_{\text{opt}}\,\underbrace{\text{cov}_{\Theta,X}(\Theta - a_{\text{opt}}\,X, X)}_{0, \text{ by } (20)} \\
&= \sigma_\Theta^2 - \frac{\text{cov}^2_{\Theta,X}(\Theta, X)}{\sigma_X^2}
\end{aligned}$$

and (15) follows. By completing the squares, it is easy to check that, for any $a \in \mathbb{R}$,

$$\mathrm{var}_{\Theta,X}(\Theta - a\,X) = \mathrm{var}_{\Theta,X}(\Theta - a\,X + a_{\mathrm{opt}}\,X - a_{\mathrm{opt}}\,X)$$

$$= \mathrm{var}_{\Theta,X}\Big((\Theta - a_{\mathrm{opt}}\,X) - (a - a_{\mathrm{opt}})\,X\Big)$$

$$= \underbrace{\mathrm{var}_{\Theta,X}(\Theta - a_{\mathrm{opt}}\,X)}_{\mathrm{MBMSE}_{\mathrm{linear}}} + (a - a_{\mathrm{opt}})^2\,\mathrm{var}_X(X)$$

$$-2\,(a - a_{\mathrm{opt}})\,\underbrace{\mathrm{cov}_{\Theta,X}(\Theta - a_{\mathrm{opt}}\,X, X)}_{0,\ \text{by (20)}}$$

$$= \underbrace{\sigma_\Theta^2 - \frac{[\mathrm{cov}_{\Theta,X}(\Theta, X)]^2}{\sigma_X^2}}_{\text{see (15)}} + \sigma_X^2\,(a - a_{\mathrm{opt}})^2 \tag{22}$$

which proves MMSE optimality of (19). $\square$

Comments:

- $$\mathrm{E}_X[\widehat{\theta}(X)] = \mathrm{E}_\Theta(\Theta)$$

  also true for the MMSE estimate, see (4).

- If $\rho_{\Theta,X} = 0$, i.e. Θ and X are *uncorrelated*, then

  $$\widehat{\theta}(X) = \mathrm{E}_\Theta(\Theta) = \mathrm{constant}$$

  i.e. LMMSE estimation *ignores* the observation X.

- If $\rho_{\Theta,X} = \pm 1$, i.e. $\Theta - \mathrm{E}_{\Theta}(\Theta)$ and $X - \mathrm{E}_{X}(X)$ are *linearly dependent* with probability one, then the LMMSE estimate is perfect.

LMMSE vs. MMSE

In general, the LMMSE estimate is *not as good* as the MMSE estimate.

Example: Suppose that

$$X \sim \mathrm{U}(-1,1) \quad \text{uniform pdf, see the table of distributions}$$

and

$$\Theta = X^2.$$

The MMSE estimate of Θ based on X is

$$\widehat{\theta}(X) = \mathrm{E}_{\Theta \,|\, X}(\Theta \,|\, X) = X^2$$

which is perfect. To find the LMMSE estimate of Θ based on X, we need

$$\begin{aligned} \mathrm{E}_X(X) &= 0 \\ \mathrm{E}_\Theta(\Theta) &= \int_{-1}^{1} x^2 \cdot \tfrac{1}{2}\, dx = \frac{1}{3} \\ \mathrm{cov}_{\Theta,X}(\Theta, X) &= \mathrm{E}_{\Theta,X}(\Theta\, X) - 0 = 0 \quad \Theta \text{ and } X \text{ uncorr.} \end{aligned}$$

yielding the LMMSE estimate

$$\widehat{\theta}(X) = \mathrm{E}_{\Theta}(\Theta) = \frac{1}{3}$$

i.e. the observation X is totally ignored even though it completely determines Θ.

An class of random signals for which the MMSE estimate is linear is the class of *jointly Gaussian random signals*, e.g. Θ and X in the additive Gaussian noise channel example on p. 8.

Linear MMSE Estimation: A Geometric Formulation

We first introduce some background:

- A *vector space* $\mathcal{V}$ (e.g. the common Euclidean space) consists of a set of vectors that is closed under two operations:
  - *vector addition*: if $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{V}$, then $\boldsymbol{v}_1 + \boldsymbol{v}_2 \in \mathcal{V}$ and
  - *scalar multiplication*: if $a \in \mathbb{R}$ and $\boldsymbol{v} \in \mathcal{V}$, then $a\,\boldsymbol{v} \in \mathcal{V}$.

- An *inner product*, (e.g. scalar product product in Euclidean spaces), is an operation $\boldsymbol{u} \cdot \boldsymbol{v}$ satisfying
  - commutativity: $\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{u}$,
  - linearity: $(a\,\boldsymbol{u} + b\,\boldsymbol{v}) \cdot \boldsymbol{w} = a\,\boldsymbol{u} \cdot \boldsymbol{w} + b\,\boldsymbol{v} \cdot \boldsymbol{w}$, and
  - the inner product of any vector with itself
    - is non-negative: $\boldsymbol{u} \cdot \boldsymbol{u} \geq 0$, and
    - $\boldsymbol{u} \cdot \boldsymbol{u} = 0$ if and only if $\boldsymbol{u} = 0$.

- The *norm* of $\boldsymbol{u}$ is defined as $\|\boldsymbol{u}\| = \sqrt{\boldsymbol{u} \cdot \boldsymbol{u}}$.

- $\boldsymbol{u}$ and $\boldsymbol{v}$ are *orthogonal* (written $\boldsymbol{u} \perp \boldsymbol{v}$) if and only if $\boldsymbol{u} \cdot \boldsymbol{v} = 0$.

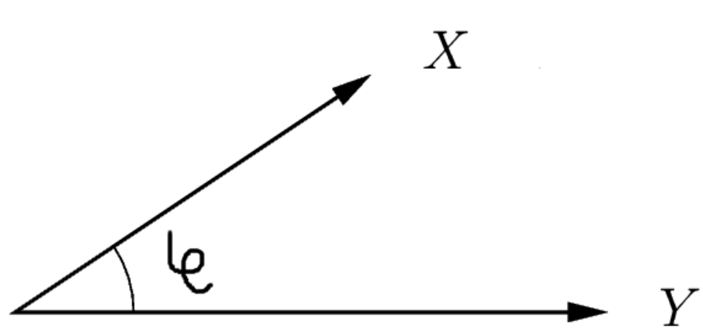
- A vector space with an inner product is called an *inner-product space*. **Example:** Euclidean space with the scalar product.

How about a vector space for random variables?

Consider random variables X and Y as vectors in an inner-product space $\mathcal{V}$ that contains all RVs defined over the same probability space, with

- vector addition: $V_1 + V_2 \in \mathcal{V}$,
- scalar multiplication: $a\,V \in \mathcal{V}$,
- inner product: $V_1 \cdot V_2 = \mathrm{cov}_{V_1,V_2}(V_1, V_2)$ (check that it is a legitimate inner product),
- the norm of V: $\|V\| = \sqrt{\mathrm{var}_V(V^2)} = \sigma_V$.

Hence,

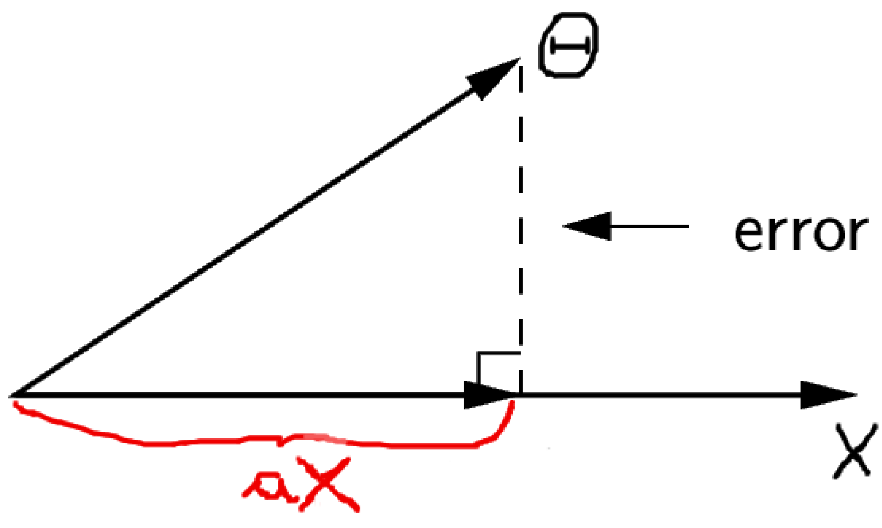


$$\begin{aligned}
\text{inner product} &\quad \Leftrightarrow \quad \operatorname{cov}_{X,Y}(X, Y) \\
\text{norm of } X &\quad \Leftrightarrow \quad \sigma_X \\
\text{norm of } Y &\quad \Leftrightarrow \quad \sigma_Y \\
\cos\varphi &\quad \Leftrightarrow \quad \rho_{X,Y}.
\end{aligned}$$

The linear MMSE estimation problem can be recast in the above geometric framework after substituting the optimal b from (16) into $\mathrm{E}_{\Theta,X}\{[\Theta - a\,X - b]^2\}$, yielding

$$\operatorname{var}_{\Theta,X}(\Theta - a\,X) = \|\Theta - a\,X\|^2.$$

We wish to minimize this variance with respect to a.



Clearly, $\|\Theta - a\,X\|^2$ is minimized if

$$(\theta - a\,X) \perp X$$

i.e. if

$$\text{cov}_{\Theta,X}(\Theta - a\,X,\; X) = 0$$

and, consequently, the MMSE-optimal linear term a is

$$a_{\text{opt}} = \frac{\text{cov}_{\Theta,X}(\Theta, X)}{\text{var}_X(X)} = \frac{\text{cov}_{\Theta,X}(\Theta, X)}{\sigma_X^2}.$$

To summarize:

Orthogonality principle:

$$(\Theta - a_{\mathrm{opt}}\, X) \perp X$$

i.e.

$$\mathrm{cov}_{\Theta,X}(\Theta - a_{\mathrm{opt}}\, X,\ X) = 0 \tag{23}$$

see (20).

Additive White Noise Channel

Consider again the communication channel example on p. 8, with input Θ having mean μ_Θ and variance τ_Θ^2 and noise W having mean zero and variance σ^2, where Θ and W are independent and the measurement X is

$$X = \Theta + W.$$

Find the LMMSE estimate of Θ based on X and the resulting BMSE ($\text{MBMSE}_{\text{linear}}$). We need

$$\begin{aligned} \mathrm{E}_\Theta(\Theta) &= \mu_\Theta \\ \mathrm{E}_X(X) &= \mathrm{E}_{X,W}(\Theta + W) = \mathrm{E}_\Theta(\Theta) + \mathrm{E}_W(W) = \mu_\Theta \end{aligned}$$

and

$$\begin{aligned} \mathrm{cov}_{\Theta,X}(\Theta, X) &= \mathrm{cov}_{\Theta,W}(\Theta, \Theta + W) \\ &\overset{\Theta \text{ and } W \text{ uncorr.}}{=} \tau_\Theta^2 \\ \mathrm{cov}_X(X) &= \mathrm{cov}_{X,W}(\Theta + W, \Theta + W) \\ &\overset{\Theta \text{ and } W \text{ uncorr.}}{=} \tau_\Theta^2 + \sigma^2. \end{aligned}$$

The LMMSE estimate of X is

$$
\begin{aligned}
\widehat{\theta}(X) &= \frac{\mathrm{cov}_{\Theta,X}(\Theta, X)}{\sigma_X^2} \cdot [X - \mathrm{E}_X(X)] + \mathrm{E}_X(X) \\
&= \frac{\tau_\Theta^2}{\tau_\Theta^2 + \sigma^2} (X - \mu_X) + \mu_X \\
&= \frac{\tau_\Theta^2}{\tau_\Theta^2 + \sigma^2} X + \frac{\sigma^2}{\tau_\Theta^2 + \sigma^2} \mu_\Theta \\
&= \frac{\frac{1}{\sigma^2} X + \frac{1}{\tau_\Theta^2} \mu_\Theta}{\frac{1}{\sigma^2} + \frac{1}{\tau_\Theta^2}}
\end{aligned}
$$

which is the same as the MMSE estimate in (8).

Example: Estimating the Bias of a Coin

Suppose that (prior) pdf of heads Θ of a coin is

$$f_\Theta(\theta) = \mathrm{U}(\theta \,|\, 0, 1) = i_{(0,1)}(\theta).$$

We flip this coin N times and record the number of heads X. Then, if the coin flips are independent, identically distributed (i.i.d.), the conditional pdf of X given $\Theta = \theta$ is

$$f_{X \,|\, \Theta}(x \,|\, \theta) = \binom{N}{x} \theta^x \, (1-\theta)^{N-x} = \mathrm{Bin}(x \,|\, N, \theta) \quad \text{binomial pdf}. \tag{24}$$

Find the MMSE and LMMSE estimates of Θ based on X.

MMSE:

$$\begin{aligned} f_{\Theta \,|\, X}(\theta \,|\, x) &\propto f_\Theta(\theta) \, f_{X \,|\, \Theta}(x \,|\, \theta) \\ &\propto i_{(0,1)}(\theta) \, \theta^x \, (1-\theta)^{N-x} \\ &= \mathrm{Beta}(\theta \,|\, x+1, N-x+1) \end{aligned}$$

see the table of distributions. Now, the MMSE estimate of Θ is

$$\widehat{\theta}_{\mathrm{MMSE}}(x) = \mathrm{E}_{\,\Theta \,|\, X}(\Theta \,|\, X = x) = \frac{x+1}{N+2}.$$

LMMSE: We need

$$\mu_\Theta = \mathrm{E}_\Theta(\Theta) = \tfrac{1}{2} \quad \text{mean of uniform}(0,1) \text{ pdf}$$

$$\mu_X = \mathrm{E}_{\Theta,X}(X) \overset{\text{iter. exp.}}{=} \mathrm{E}_\Theta[\mathrm{E}_{X\,|\,\Theta}(X\,|\,\Theta)]$$

$$= \mathrm{E}_\Theta(\underbrace{N\,\Theta}_{\text{mean of binomial pdf in (24)}}) = \tfrac{1}{2}N$$

and

$$\sigma_X^2 \overset{\text{cond. var.}}{=} \mathrm{E}_\Theta\{\underbrace{\mathrm{var}_{X\,|\,\Theta}(X\,|\,\Theta)}_{\text{var of binomial in (24)}}\} + \mathrm{var}_\Theta\{\underbrace{\mathrm{E}_{X\,|\,\Theta}[X\,|\,\Theta]}_{\text{mean of binomial in (24)}}\}$$

$$= \mathrm{E}_\Theta[N\,\Theta\,(1-\Theta)] + \mathrm{var}_\Theta(N\,\Theta)$$

$$= N\,\mathrm{E}_\Theta[\Theta\,(1-\Theta)] + N^2\,\mathrm{var}_\Theta(\Theta)$$

$$= N\,(\frac{1}{2} - \frac{1}{3}) + N^2\,\frac{1}{12} = \frac{N\,(N+2)}{12}$$

$$\mathrm{cov}_{\Theta,X}(\Theta, X) = \mathrm{E}_{\Theta,X}(\Theta\,X) - \mu_\Theta\,\mu_X$$

$$\overset{\text{iterated exp.}}{=} \mathrm{E}_\Theta\{\mathrm{E}_{X\,|\,\Theta}[\Theta\,X\,|\,\Theta]\} - \frac{N}{4}$$

$$= \mathrm{E}_\Theta[\Theta\,\mathrm{E}_{X\,|\,\Theta}(X\,|\,\Theta)] - \frac{N}{4} = \mathrm{E}_\Theta\{\Theta\,\underbrace{N\,\Theta}_{\text{mean of binomial in (24)}}\} - \frac{N}{4}$$

$$= \frac{N}{3} - \frac{N}{4} = \frac{N}{12}.$$

Now,

$$\begin{aligned}
\widehat{\theta}(X) &= \frac{\mathrm{cov}_{\Theta,X}(\Theta, X)}{\sigma_X^2} \cdot (X - \mu_X) + \mu_\Theta \\
&= \frac{N/12}{N\,(N+2)/12} \cdot (X - \tfrac{1}{2}\,N) + \tfrac{1}{2} = \frac{X+1}{N+2}.
\end{aligned}$$

In this example, the MMSE and LMMSE estimates of θ are the same.

Linear MMSE Estimation: the Vector Case (FIR Wiener Filter)

Consider the signal of interest with prior knowledge described by the pdf

$$\Theta \sim f_{\Theta}(\theta)$$

and an N-dimensional random vector $\boldsymbol{X}$ representing the observations.

- The MMSE estimate of X is the conditional expectation

$$\mathrm{E}_{\Theta \,|\, \boldsymbol{X}}(\Theta \,|\, \boldsymbol{X})$$

  which may be difficult to find in practice, since it requires knowledge of the joint distribution of Θ and $\boldsymbol{X}$.

- The linear MMSE estimate of X is easier to find, since it depends only on the means, variances, and covariances of the random variables and vectors involved.

Linear MMSE Estimation via the Orthogonality Principle

We wish to find an $N \times 1$ vector $\boldsymbol{a}$ and a constant b such that

$$\widehat{\theta}(\boldsymbol{X}) = \boldsymbol{a}^T \boldsymbol{X} + b = \sum_{i=1}^{N} a_i X_i + b$$

minimizes the BMSE

$$\text{BMSE} = \text{E}_{\Theta,\boldsymbol{X}}\{[\Theta - \widehat{\theta}(\boldsymbol{X})]^2\}$$

where

$$\boldsymbol{X} = \begin{bmatrix} X[0] \\ \dots \\ X[N-1] \end{bmatrix}.$$

Suppose first that the constant vector $\boldsymbol{a}$ has already been chosen. Then, choosing the constant b to minimize the BMSE

$$\text{BMSE} = \text{E}_{\Theta,X}[(\Theta - \boldsymbol{a}^T \boldsymbol{X} - b)^2]$$

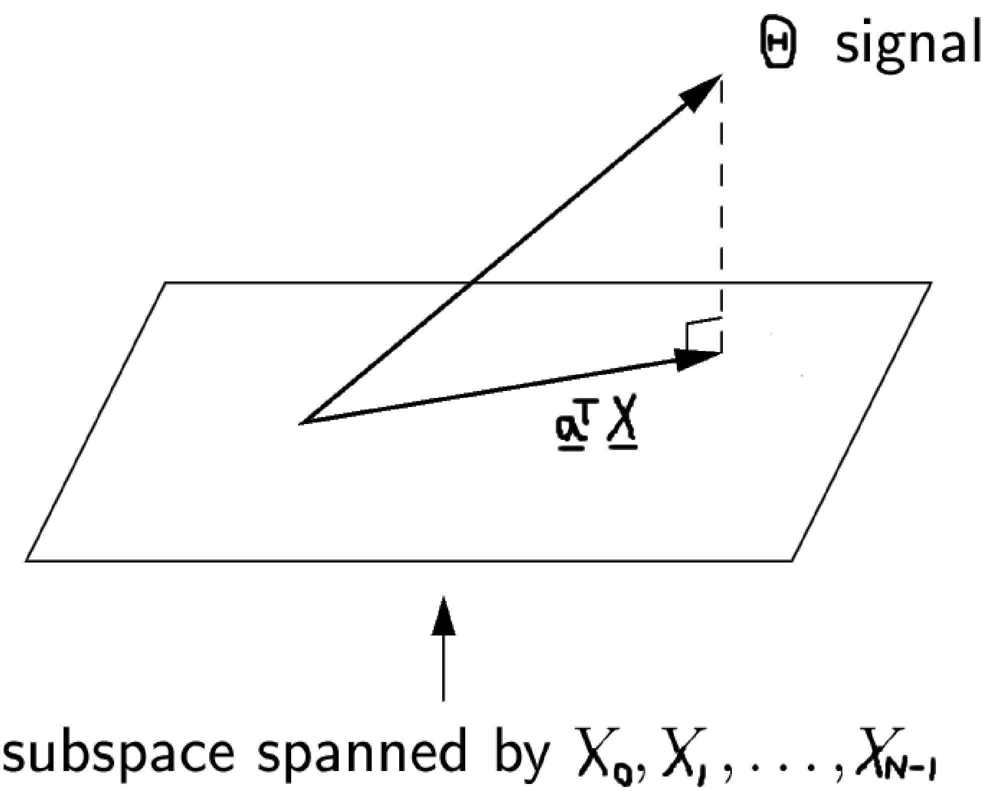
is equivalent to finding b that minimizes $\text{E}_{\Xi}[(\Xi - b)^2]$ where $\Xi = \Theta - \boldsymbol{a}^T \boldsymbol{X}$. This problem is solved in Lemma 1 and the optimal b is

$$b = \text{E}_{\Xi}(\Xi) = \text{E}_{\Theta,X}(\Theta - \boldsymbol{a}^T \boldsymbol{X}) = \text{E}_{\Theta}(\Theta) - \boldsymbol{a}^T \text{E}_{\boldsymbol{X}}(\boldsymbol{X}). \quad (25)$$

We view $\Theta, X[0], \ldots, X[N-1]$ as vectors in an inner-product space. The linear MMSE estimation problem can be cast into our geometric framework after substituting the optimal b in (25) into $\mathrm{BMSE} = \mathrm{E}_{\Theta,\boldsymbol{X}}\{[\Theta - \widehat{\theta}(\boldsymbol{X})]^2\}$, yielding

$$\mathrm{var}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}^T\boldsymbol{X}) = \|\Theta - \boldsymbol{a}^T\boldsymbol{X}\|^2. \qquad (26)$$

We minimize this variance with respect to $\boldsymbol{a}$.



Clearly, $\|\Theta - \boldsymbol{a}^T\boldsymbol{X}\|^2$ is minimized if $\boldsymbol{a}$ is chosen to satisfy the *orthogonality principle*:

$(\Theta - \boldsymbol{a}^T\boldsymbol{X}) \perp$ subspace $\mathcal{V}_N$ spanned by $X[0], X[1], \ldots, X[N-1]$

or, equivalently,

$$\mathrm{cov}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}^T \boldsymbol{X}, X[n]) = 0 \quad n = 0, 1, \ldots, N-1 \qquad (27)$$

which gives the following set of equations:

$$\mathrm{cov}_{\Theta,X[n]}(\Theta, X[n]) - \mathrm{cov}_{\boldsymbol{X}}\Big(\sum_{l=0}^{N-1} a_l\, X[l],\, X[n]\Big) = 0$$

or

$$\sum_{l=0}^{N-1} \mathrm{cov}_{X[n],X[l]}(X[n], X[l])\, a_l = \mathrm{cov}_{X[n],\Theta}(X[n], \Theta). \qquad (28)$$

Define the crosscovariance vector between $\boldsymbol{X}$ and Θ and covariance matrix of $\boldsymbol{X}$ as

$$\boldsymbol{\sigma}_{\boldsymbol{X},\Theta} = \mathrm{cov}_{\boldsymbol{X},\Theta}(\boldsymbol{X}, \Theta) = \begin{bmatrix} \mathrm{cov}_{X[0],X}(X[0], \Theta) \\ \mathrm{cov}_{X[1],\Theta}(X[1], \Theta) \\ \vdots \\ \mathrm{cov}_{X[N-1],\Theta}(X[N-1], \Theta) \end{bmatrix}$$

and

$$\Sigma_{\boldsymbol{X}} = \mathrm{cov}_{\boldsymbol{X}}(\boldsymbol{X})$$

and use these definitions to compactly write (28):

$$\Sigma_{\boldsymbol{X}}\, \boldsymbol{a} = \boldsymbol{\sigma}_{\boldsymbol{X},\Theta}.$$

If $\Sigma_{\boldsymbol{X}}$ is a positive definite matrix, we can solve for $\boldsymbol{a}$:

$$\boldsymbol{a}_{\text{opt}} = \Sigma_{\boldsymbol{X}}^{-1}\,\boldsymbol{\sigma}_{\boldsymbol{X},\Theta} \tag{29}$$

and, finally, the LMMSE estimate of Θ is [using (25)]

$$\begin{aligned} \widehat{\theta}(\boldsymbol{X}) &= \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X} + \mathrm{E}_{\Theta}(\Theta) - \boldsymbol{a}_{\text{opt}}^T\,\mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X}) \\ &= \underbrace{\boldsymbol{\sigma}_{\boldsymbol{X},\Theta}^T \Sigma_{\boldsymbol{X}}^{-1}}_{\boldsymbol{a}_{\text{opt}}^T}\,[\boldsymbol{X} - \mathrm{E}_{\boldsymbol{X}}(\boldsymbol{X})] + \mathrm{E}_{\Theta}(\Theta). \end{aligned} \tag{30}$$

Compare this result to the scalar case in (13):

$$\widehat{\theta}(X) = \underbrace{\frac{\mathrm{cov}_{\Theta,X}(\Theta, X)}{\sigma_X^2}}_{a_{\text{opt}}} \cdot [X - \mathrm{E}_X(X)] + \mathrm{E}_{\Theta}(\Theta).$$

We now find the minimum BMSE of our LMMSE estimator: substitute (29) into (26), yielding

$$\begin{aligned} \mathrm{MBMSE}_{\text{linear}} &= \mathrm{cov}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}, \Theta - \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}) \\ &= \mathrm{cov}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}, \Theta) - \mathrm{cov}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}, \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}) \\ &= \mathrm{cov}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}, \Theta) - \underbrace{\mathrm{cov}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}, \boldsymbol{X})}_{= \boldsymbol{0}, \text{ see (27)}}\,\boldsymbol{a}_{\text{opt}} \\ &= \mathrm{cov}_{\Theta,\boldsymbol{X}}(\Theta - \boldsymbol{a}_{\text{opt}}^T\,\boldsymbol{X}, \Theta) \end{aligned} \tag{31}$$

which can also be written as

$$\text{MBMSE}_{\text{linear}} = \text{cov}_{\Theta,\boldsymbol{X}}(\Theta - \widehat{\theta}(\boldsymbol{X}), \Theta) \tag{32}$$

and further simplified:

$$\begin{aligned} \text{MBMSE}_{\text{linear}} &= \sigma_X^2 - \boldsymbol{a}_{\text{opt}}^T \underbrace{\text{cov}_{\boldsymbol{X},\Theta}(\boldsymbol{X}, \Theta)}_{\boldsymbol{\sigma}_{\boldsymbol{X},\Theta}} \\ &\overset{\text{see (29)}}{=} \sigma_\Theta^2 - \boldsymbol{\sigma}_{\boldsymbol{X},\Theta}^T \, \Sigma_{\boldsymbol{X}}^{-1} \, \boldsymbol{\sigma}_{\boldsymbol{X},\Theta}. \end{aligned} \tag{33}$$

Compare this result to the scalar case in (15):

$$\text{MBMSE}_{\text{linear}} = \sigma_\Theta^2 - \frac{\text{cov}_{\Theta,X}^2(\Theta, X)}{\sigma_X^2}.$$

Example: Additive Noise Channel

Again:

$$\Theta \sim \mathcal{N}(\mu_\Theta, \tau_\Theta^2)$$

where μ_Θ and τ_Θ^2 are known hyperparameters. We collect multiple observations $X[n]$, modeled as

$$X[n] = \Theta + W[n] \quad n = 0, 1, \ldots, N-1$$

where $W[n]$ are zero-mean uncorrelated RVs with known variance σ^2. We also know that Θ and $W[n]$ are uncorrelated for all n. Find the LMMSE estimate of Θ based on

$$\boldsymbol{X} = \begin{bmatrix} X[0] \\ \ldots \\ X[N-1] \end{bmatrix}.$$

Find also the minimum BMSE.

By the orthogonality principle (27), we have:

$$\mathrm{cov}_{\Theta, X[n]}(\Theta, X[n]) - \mathrm{cov}_{\boldsymbol{X}}\Big(\sum_{l=0}^{N-1} a_l\, X[l],\, X[n]\Big) = 0$$

for $n = 0, 1, \ldots, N-1$. Here,

$$\begin{aligned} \mathrm{cov}_{\Theta, X[n]}(\Theta, X[n]) &= \mathrm{cov}_{\Theta, W[n]}(\Theta + W[n], \Theta) \\ &= \mathrm{cov}_{\Theta}(\Theta, \Theta) + \mathrm{cov}_{W[n], X}(W[n], \Theta) \\ &= \mathrm{var}_{\Theta}(\Theta) = \tau_{\Theta}^2 \qquad (34) \\ \mathrm{cov}_{X[l], X[n]}(X[l], X[n]) &= \mathrm{cov}_{\Theta, W}(\Theta + W[l], \Theta + W[n]) \\ &= \begin{cases} \tau_{\Theta}^2, & l \neq n \\ \tau_{\Theta^2} + \sigma^2, & l = n \end{cases} \qquad (35) \end{aligned}$$

and, therefore,

$$\begin{aligned} \tau_{\Theta}^2 &= (\tau_{\Theta^2} + \sigma^2)\, a_0 + \tau_{\Theta}^2\, a_1 + \ldots + \tau_{\Theta}^2\, a_{N-1} \\ \tau_{\Theta}^2 &= \tau_{\Theta}^2\, a_0 + (\tau_{\Theta^2} + \sigma^2)\; a_1 + \ldots + \tau_{\Theta}^2\, a_{N-1} \\ \cdots \\ \tau_{\Theta}^2 &= \tau_{\Theta}^2\, a_0 + \tau_{\Theta}^2\, a_1 + \ldots + (\tau_{\Theta^2} + \sigma^2)\, a_{N-1}. \end{aligned}$$

Now, by symmetry,

$$a_{\mathrm{opt},1} = a_{\mathrm{opt},2} = \ldots = a_{\mathrm{opt},N} = \frac{\tau_{\Theta}^2}{N\,\tau_{\Theta}^2 + \sigma^2}$$

yielding

$$\begin{aligned}
\widehat{\theta}(\boldsymbol{X}) &= \frac{\tau_\Theta^2}{N\,\tau_\Theta^2 + \sigma^2} \sum_{n=0}^{N-1} (X[n] - \mu_\Theta) + \mu_\Theta \\
&= \frac{\tau_\Theta^2}{N\,\tau_\Theta^2 + \sigma^2} \Big(\sum_{n=0}^{N-1} X[n] \Big) + \frac{\sigma^2}{N\,\tau_\Theta^2 + \sigma^2}\, \mu_\Theta \qquad (36) \\
&= \frac{\frac{N}{\sigma^2}\,\overline{X} + \frac{1}{\tau_\Theta^2}\,\mu_\Theta}{\frac{N}{\sigma^2} + \frac{1}{\tau_\Theta^2}}
\end{aligned}$$

where

$$\overline{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n].$$

The minimum average MSE of our LMMSE estimator follows by using (31):

$$
\begin{aligned}
\text{MBMSE}_{\text{linear}} &= \text{cov}_{\Theta,\boldsymbol{X}}(\Theta - \widehat{\theta}(\boldsymbol{X}), \Theta) \\
&= \tau_\Theta^2 - \text{cov}_{\Theta,\boldsymbol{X}}(\widehat{\theta}(\boldsymbol{X}), \Theta) \\
&\overset{\text{see (36)}}{=} \tau_\Theta^2 - \text{cov}_{\Theta,\boldsymbol{X}}\left(\frac{\tau_\Theta^2}{N\,\tau_\Theta^2 + \sigma^2}\Big(\sum_{n=0}^{N-1} X[n]\Big), \Theta\right) \\
&\overset{\text{see (34)}}{=} \tau_\Theta^2 - \frac{\tau_\Theta^2}{N\,\tau_\Theta^2 + \sigma^2} \sum_{n=0}^{N-1} \text{cov}_{X[n],\Theta}(X[n], \Theta) \\
&= \tau_\Theta^2 - \frac{\tau_\Theta^2}{N\,\tau_\Theta^2 + \sigma^2} N\,\tau_\Theta^2 \\
&= \frac{\tau_\Theta^2\,\sigma^2}{N\,\tau_\Theta^2 + \sigma^2} \\
&= \left(\frac{N}{\sigma^2} + \frac{1}{\tau_\Theta^2}\right)^{-1}
\end{aligned}
$$

which is the same as τ_N^2 in (15) of handout # 4.