Lecture 9: Minimum Mean-Square Estimation
Definition 39. A set $G$ with a commutative operation $+: G \times G \rightarrow G$ is an additive group if

1. it is closed under $+$ (i.e., $a + b \in G$ for all $a, b \in G$).

2. it has an additive identity (i.e., there exists an element $0 \in G$ such that $0 + a = a$ for all $a \in G$).

3. Every element has an additive inverse (i.e., for all $a \in G$ there exist an element $-a$ such that $a + (-a) = 0$).
**Definition 40.** A vector space $V$ over $\mathbb{R}$ is an additive group such that

1. $xv \in V$ for all $v \in V$ and $x \in \mathbb{R}$.

2. $0v = 0$ for all $v \in V$.

3. $1v = v$ for all $v \in V$.  

\[\diamondsuit\]
Definition 41. A norm is a function $\| \cdot \| : V \to \mathbb{R}_+$ that satisfies the following properties:

1. $\| v \| = 0$ if and only if $v = 0$.

2. $\| v + u \| \leq \| v \| + \| u \|$ (triangle inequality).

3. $\| x v \| = |x| \| v \|$ for all $v \in V$ and $x \in \mathbb{R}$. 

\[\Box\]
**Definition 42.** A normed vector space is a vector space $V$ with a norm $\| \cdot \|$. ◊

Notice: a norm is a “distance” function.

- For example, one can check that the norm defined as

$$\|v\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$$

where $V = \mathbb{R}^n$ is the standard Euclidean $n$-dimensional vector space over $\mathbb{R}$, defines a distance in the usual sense (length of the vector joining two points in $\mathbb{R}^n$).

- Let $v, u \in \mathbb{R}^n$, then

$$\|v - u\|_2 = \sqrt{\sum_{i=1}^{n} (v_i - u_i)^2}$$

is the Euclidean distance between the points (vectors) $v$ and $u$. 

Copyright G. Caire (Sample Lectures)
Definition 43. Given a vector space $V$ over $\mathbb{R}$, an inner product is a function $(\cdot, \cdot) : V \times V \to \mathbb{R}$ with the following properties:

1. $(v, u) = (u, v)$ (symmetry).

2. $(xv, u) = x(v, u)$, for all $v, u \in V$ and $x \in \mathbb{R}$ (scaling).

3. $(v_1 + v_2, u) = (v_1, u) + (v_2, u)$ (linearity).

4. $(v, v) \geq 0$, with equality if and only if $v = 0$.

A vector space with an inner product is called inner product space.
Theorem 41. *Cauchy-Schwarz inequality:*

\[(v, u)^2 \leq (v, v)(u, u)\]

with equality if and only if \(a v = b u\), with \(a, b \in \mathbb{R}\) not both zero.

Theorem 42. Let \(V\) be an inner product space. Then, the following is a norm (called 2-norm, or standard Euclidean norm):

\[\|v\|_2 = \sqrt{(v, v)}\]
Least Squares approximation

Consider the following problem. Let $x$ denote a point (vector) in some vector space $V$ over $\mathbb{R}$ and let $y_1, \ldots, y_m$ be a given collection of vectors: we wish to find the “best” approximation of $x$ by a linear combination of the vectors $\{y_i\}$.

We have to give a rigorous meaning to the term “best”: if $V$ is an inner product space, we shall consider the minimum distance approximation, that is, we look for

$$\hat{x} = \sum_{i=1}^{m} y_i a_i$$

such that

$$\|x - \hat{x}\|_2^2 = (x - \hat{x}, x - \hat{x})$$

is minimum.

This approximation is called (linear) “Least-Squares” (some people call it “linear regression”).
• A brute-force approach: we can write

\[
\| x - \hat{x} \|_2^2 = \| x \|_2^2 - 2(x, \hat{x}) + \| \hat{x} \|_2^2 \\
= \| x \|_2^2 - 2 \sum_{i=1}^{m} (x, y_i) a_i + \sum_{i=1}^{m} \sum_{j=1}^{m} a_i (y_i, y_j) a_j \\
= \| x \|_2^2 - 2 r_{xy}^T a + a^T G_y a
\]

where we define the “cross-correlation vector”

\[
r_{xy} = ((x, y_1), \ldots, (x, y_m))^T
\]

and the matrix of inner products (Gram matrix)

\[
G_y = \begin{bmatrix}
(y_1, y_1) & (y_1, y_2) & \cdots & (y_1, y_m) \\
(y_2, y_1) & (y_2, y_2) & \cdots & \vdots \\
& \ddots & \ddots & \vdots \\
(y_m, y_1) & (y_m, y_2) & \cdots & (y_m, y_m)
\end{bmatrix}
\]
• Notice that $G_y$ is symmetric and positive semidefinite.

• Taking the gradient of the distance function with respect to $a$, we obtain the equation

$$G_y a = r_{xy}$$

• Assuming for simplicity that $G_y$ is invertible (otherwise, we can eliminate some linearly dependent $y_i$ and obtain the same subspace), we obtain $a = G_y^{-1} r_{xy}$.

• **Observation:** notice that the solution $\hat{x}$ satisfies the following orthogonality condition:

$$(x - \hat{x}, y_i) = 0, \quad \forall \ i = 1, \ldots, m$$
Linear Minimum Mean-Square Error estimation

- We have two jointly distributed random vectors $X$ and $Y$.

- We observe $Y$ and we wish to “guess” the value of $X$ in some optimal sense.

- Analogously to what done before, we define the following error function: Mean-Square-Error (MSE)

$$\text{mse} = E \left[ \left\| X - \hat{X} \right\|_2^2 \right]$$

- We seek an estimator $\hat{X}$ in the form of an affine function of the observation $Y$, that is,

$$\hat{X} = AY + b$$
First, notice that for any mean vectors $\mathbf{m}_x$ and $\mathbf{m}_y$ and any estimator $\hat{\mathbf{X}}$, we can always reduce the problem to a zero-mean case by considering $\mathbf{X}' = \mathbf{X} - \mathbf{m}_x$, $\mathbf{Y}' = \mathbf{Y} - \mathbf{m}_y$. If $\hat{\mathbf{X}}'$ is the minimum MSE (MMSE) estimator for $\mathbf{X}'$ given $\mathbf{Y}'$, then

$$\hat{\mathbf{X}} = \mathbf{m}_x + \hat{\mathbf{X}}'$$

is the optimal estimator for $\mathbf{X}$ given $\mathbf{Y}$.

We restrict to the zero-mean case, and seek $\hat{\mathbf{X}}$ in the form $\mathbf{A}\mathbf{Y}$. The orthogonality principle yields the condition

$$(\mathbf{X} - \hat{\mathbf{X}}, \mathbf{Y}) = \mathbb{E} \left[ (\mathbf{X} - \hat{\mathbf{X}})^\intercal \mathbf{Y} \right] = \text{tr} \left( \mathbb{E} \left[ (\mathbf{X} - \hat{\mathbf{X}})\mathbf{Y}^\intercal \right] \right) = 0$$

that, explicitly, writes

$$\text{tr} \left( \mathbb{E} \left[ \mathbf{X}\mathbf{Y}^\intercal \right] - \mathbf{A}\mathbb{E} \left[ \mathbf{Y}\mathbf{Y}^\intercal \right] \right) = 0$$
We can solve for $A$ as follows:

$$A\mathbb{E}[YY^T] = \mathbb{E}[XY^T] \Rightarrow A = \mathbb{E}[XY^T] (\mathbb{E}[YY^T])^{-1}$$

In conclusions, in the general case (non-zero mean), we let

$$\Sigma_{xy} = \text{cov}(X, Y) = \mathbb{E}[(X - m_x)(Y - m_y)^T], \quad \Sigma_y = \text{cov}(Y)$$

and we obtain the linear MMSE estimator (Wiener filter) of $X$ from $Y$ as

$$\hat{X} = m_x + \Sigma_{xy} \Sigma_y^{-1} (Y - m_y)$$
The error covariance matrix is given by

\[
\text{cov}(\mathbf{X} - \hat{\mathbf{X}}) = \mathbb{E} \left[ (\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})^\top \right]
\]

\[
= \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}
\]

The total MMSE, is given by \(\mathbb{E}[\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2] = \text{tr}(\text{cov}(\mathbf{X} - \hat{\mathbf{X}}))\).

**Notice:** The estimation error vector \(\mathbf{X} - \hat{\mathbf{X}}\) is uncorrelated with the observation vector \(\mathbf{Y}\).
• With the same setting as before, we now seek an estimator \( \hat{X} = g(Y) \), in the space of all (measurable) functions of the observation \( Y \).

**Theorem 43.** The MMSE estimator of \( X \) given \( Y \) is the conditional mean

\[
\hat{X} = \mathbb{E}[X|Y]
\]

**Proof.**
We use the orthogonality principle: the optimal estimator \( \hat{X} \) must satisfy

\[
\mathbb{E} \left[ (X - \hat{X})^T g(Y) \right] = 0, \quad \text{for all functions } g
\]
Let’s check with the conditional mean:

\[
\mathbb{E} \left[ (\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^T g(\mathbf{Y}) \right] = \mathbb{E} \left[ \mathbb{E} \left[ (\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^T g(\mathbf{Y}) | \mathbf{Y} \right] \right] \\
= \mathbb{E} \left[ \mathbb{E} \left[ \mathbf{X}^T g(\mathbf{Y}) | \mathbf{Y} \right] - \mathbb{E}[\mathbf{X}|\mathbf{Y}]^T g(\mathbf{Y}) \right] \\
= \mathbb{E} \left[ \mathbb{E} \left[ \mathbf{X}|\mathbf{Y} \right]^T g(\mathbf{Y}) - \mathbb{E}[\mathbf{X}|\mathbf{Y}]^T g(\mathbf{Y}) \right] \\
= 0
\]
The Gaussian case

- If $X, Y$ are jointly Gaussian, then the linear MMSE estimator and the optimal MMSE estimator coincide.

- In order to see this, recall Theorem 14

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_{x|y})}} \exp \left( -\frac{1}{2} (x - m_{x|y})^\top \Sigma^{-1}_{x|y} (x - m_{x|y}) \right)$$

where the conditional mean value is given by

$$m_{x|y} = \mathbb{E}[X|Y = y] = m_x + \Sigma_{xy} \Sigma^{-1}_y (y - m_y)$$

and the conditional covariance matrix is given by

$$\Sigma_{x|y} = \mathbb{E}[(X - m_{x|y})(X - m_{x|y})^\top|Y = y] = \Sigma_x - \Sigma_{xy} \Sigma^{-1}_y \Sigma_{yx}$$
• Hence, the (general) MMSE estimator of $X$ given $Y$ coincides with the linear MMSE estimator (Wiener filter) in the Gaussian case:

$$\hat{X} = \mathbb{E}[X|Y] = m_x + \Sigma_{xy}\Sigma_y^{-1}(Y - m_y)$$

• MMSE decomposition:

$$X = \hat{X} + (X - \hat{X}) = \hat{X} + V$$

where the MMSE estimator $\hat{X}$ and the estimation error vector $V$ are independent,

$$\hat{X} \sim \mathcal{N}(m_x, \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}), \quad V \sim \mathcal{N}(0, \Sigma_{x|y})$$
End of Lecture 9