Mean Square Estimation
Waltenegus Dargie
Chair for Computer Networks

• A random vector is a vector $X = [X_1, X_2, \ldots, X_n]$ whose components $X_i$ are random variables.

• The probability that $X$ is in a region $D$ of the $n$-dimensional space equals the probability mass on $D$

$$P\{X \in D\} = \int_D f(X) dX \quad X = [X_1, X_2, \ldots, X_n]$$

where

$$f(X) = f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}$$

is the joint (or, multivariate) density of the random variables $X_i$. 
The joint distribution is given by

\[ F(X) = F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \]

If we substitute in \( F(X) = F(x_1, x_2, \ldots, x_n) \) certain variables by \( \infty \), we obtain the joint distribution of the remaining variables.

If we integrate \( f(X) = f(x_1, x_2, \ldots, x_n) \) with respect to certain variables, we obtain the joint density of the remaining variables.

For example:

\[ F(x_1, x_3) = F(x_1, \infty, x_3, \infty) \quad f(x_1, x_3) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4) dx_2 \, dx_4 \]
• Given \( k \) functions
\[
g_1(X), g_2(X), \ldots, g_k(X)
\]
we form the random variables
\[
Y_1 = g_1(X), Y_2 = g_2(X), \ldots, Y_k = g_k(X)
\]

• The statistics of these random variables can be determined in terms of \( X \).

• To find the density \( f_y(y_1, y_2, \ldots, y_n) \) of the random variables \( Y = [Y_1, Y_2, \ldots, Y_n] \) for a specific set of numbers \( y_1, y_2, \ldots, y_n \), we solve the system
\[
y_1 = g_1(X), y_2 = g_2(X), \ldots, y_k = g_k(X)
\]
• If this system has no solutions, then

\[ f_y(y_1, y_2, ..., y_n) = 0. \]

• If it has a single solution: \( X = [x_1, x_2, \cdots x_n] \), then

\[ f_y(y_1, y_2, ..., y_n) = \frac{f_y(x_1, x_2, ..., x_n)}{|J(x_1, x_2, ..., x_n)|} \]

where the jacobian of the transformation is given by

\[
J(x_1, x_2, ..., x_n) = \left| \begin{array}{cccc}
\frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\
\frac{\partial g_2}{\partial x_1} & \cdots & \frac{\partial g_2}{\partial x_n} \\
\cdots & \cdots & \cdots \\
\frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \\
\end{array} \right|
\]
• Suppose that $S = S_1 \times \ldots \times S_n$ is a combined experiment and the random variables $X_i$ depend only on the outcome of $\xi_i$ of $S_i$: 

$$X_i(\xi_1 \ldots \xi_i \ldots \xi_n) = X_i(\xi_i) \quad i = 1, \ldots, n$$

• If the experiment $S_i$ are independent, then the random variables $X_i$ are independent as well.

• Now suppose that $X$ is a random variable defined on an experiment $S^n = S \times \ldots \times S$ and we define the random experiment $X_i$ such that 

$$X_i(\xi_1 \ldots \xi_i \ldots \xi_n) = X_i(\xi_i) \quad i = 1, \ldots, n$$
From this it follows that the distribution $F_i(x_i)$ of $X_i$ equals the distribution $F_x(x)$ of the random variable $X$.

Thus, if an experiment is performed $n$ times, the random variables $X_i$ are independent and they have the same distribution.

These random variables are called i.i.d. – independent, identically distributed.
Suppose we measure an object of length $\eta$ with $n$ instruments of varying accuracies. The results of the measurement are $n$ random variables:

$$X_i = \eta + \nu_i \quad E\{\nu_i\} = 0 \quad E\{\nu_i^2\} = \sigma^2$$

Where $\nu_i$ are the measurement errors which we assume are independent and with zero mean.

Our aim is to determine the unbiased, minimum variance, linear estimation of $\eta$. 
Measurement Errors

• We wish to find \( n \) constant \( \alpha_i \) such that
  \[
  \hat{\eta} = \alpha_1 X_1 + ... + \alpha_n X_n
  \]
is a random variable with mean
  \[
  E\{\hat{\eta}\} = \alpha_1 E\{X_1\} + ... + \alpha_n E\{X_n\} = \eta
  \]
and its variance \( V = \alpha_1^2 \sigma_1^2 + ... + \alpha_n^2 \sigma_n^2 \) is minimum.

• Our problem is to minimize the variance subject to the constraint:
  \[
  \alpha_1 + ... + \alpha_n = 1
  \]
Measurement Errors

• For any $\lambda$, (Lagrange multiplier):

\[ V = \alpha_1^2 \sigma_1^2 + ... + \alpha_n^2 \sigma_n^2 - \lambda(\alpha_1 + ... + \alpha_n - 1) \]

• Hence, $V$ is minimum if

\[ \frac{\partial V}{\partial \alpha_i} = 2\alpha_i \sigma_i^2 - \lambda = 0 \quad \alpha_i = \frac{\lambda}{2\sigma_i^2} \]

• Solving for $\lambda$ and $\eta$ results

\[ \frac{\lambda}{2} = V = \frac{1}{1/\sigma_1^2 + ... + 1/\sigma_n^2} \]

\[ \hat{\eta} = \frac{x_1/\sigma_1^2 + ... + x_n/\sigma_n^2}{1/\sigma_1^2 + ... + 1/\sigma_n^2} \]
• The covariance $C_{ij}$ of two random variables $X_i$ and $X_j$ is defined as:

$$C_{ij} = E\left\{(X_i - \eta_i)(X_j^* - \eta_j^*)\right\}$$

• The variance of $X_i$ is:

$$C_{ii} = E\left\{(X_i - \eta_i)(X_i^* - \eta_i^*)\right\} = \sigma_i^2 = E\left\{|X_i^2|\right\} - |E\{X_i\}|^2$$

• Two random variables are said to be (mutually) uncorrelated if their covariance is zero, in which case if $X = X_1 + ... + X_n$, then $\sigma_x^2 = \sigma_1^2 + ... + \sigma_n^2$
Mean Square Estimation

• A correlation matrix describes the correlation between random variables

\[
R_n = \begin{bmatrix} R_{11} & \ldots & R_{1n} \\ \vdots & \ddots & \vdots \\ R_{n1} & \ldots & R_{nn} \end{bmatrix}
\]

where \( R_{ij} = E(X_i X_j^*) = R_{ji}^* \)

• Likewise, the covariance matrix is given as

\[
C_n = \begin{bmatrix} C_{11} & \ldots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \ldots & C_{nn} \end{bmatrix}
\]

• Thus \( C_{ij} = R_{ij} - \eta_i \eta_j^* \)
• Suppose we wish to estimate the random variable $Y$ in terms of another random variable $X$.

• The optimality criterion is the minimization of the mean square value (MS) of the estimation error.

• As we know, the distribution function $F(y)$ of the random variable $Y$ determines completely its statistics.

• This does not mean, however, knowledge of $F(y)$ is sufficient to predict the value of $Y(\xi)$ at some future trial.
Mean Square Estimation

• Suppose, however, that we wish to estimate the unknown $Y(\xi)$ by some constant $c$.

• If $Y$ is estimated by a constant $c$, then, at a particular trial the error $Y(\xi) - c$ results and our goal must be to select $c$ so as to minimise the error in some sense.

• One option is to select $c$ such that the mean square error, $\{Y(\xi) - c\}^2$ should be minimum.

$$e = E\{ (y - c)^2 \} = \int_{-\infty}^{\infty} (y - c)^2 f(y) \, dy$$
Mean Square Estimation

• Clearly $e$ depends on $c$, and it is minimum if,

\[
\frac{de}{dc} = 2 \int_{-\infty}^{\infty} (y - c)f(y) \, dy = 0
\]

\[
c = \int_{-\infty}^{\infty} yf(y) \, dy = E\{Y\}.
\]
• Suppose we wish to estimate \( Y \) not by a constant but by a function \( c(X) \) of the random variable \( X \).

• Our assignment now is to find the function \( c(X) \) that minimizes the MS.

\[
e = E\{ [Y - c(X)]^2 \} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [Y - c(X)]^2 f(x, y) \, dx \, dy
\]

• Since \( f(x, y) = f(y \mid x) \, f(x) \)

\[
e = \int_{-\infty}^{\infty} f(x) \int_{-\infty}^{\infty} [Y - c(X)]^2 \, f(y \mid x) \, dy \, dx
\]
Nonlinear MS Estimation

• The last integrands are positive, and hence, $e$ will be minimum if the inner integral is minimum for every $x$.

• The integral is of the form we previously derived for $c$; therefore, if $c$ is replaced by $c(X)$ and $f(y)$ by $f(y|x)$:

$$c(x) = \int_{-\infty}^{\infty} y f(y \mid x) \, dy = E\{Y \mid x\}.$$ 

• Non-linear MS estimation is the best estimation but the computation is usually intractable.
Linear MS Estimation

- The linear estimation problem is the estimation of the random variable $Y$ in terms of a linear function $Y = AX + B$.
- The assignment now is to find the constants $A$ and $B$ so as to minimise the MS error
  
  $$e = E \left\{ \left[ Y - (AX + B) \right]^2 \right\}$$

- We maintain that $e = e_m$ is minimum if
  
  $$A = \frac{r \sigma_y}{\sigma_x} \quad B = \eta_y - A \eta_x$$
Linear MS Estimation

• For a given $A$, $e$ is the MS error of the estimation of $Y - AX$ by the constant $B$. Therefore,

$$B = E\{Y - AX\} = E\{Y\} - AE\{X\} = \eta_y - A\eta_x$$

$$e = E \left\{ \left[ (Y - \eta_y) - A(X - \eta_x) \right]^2 \right\} = \sigma_y^2 - 2Ar\sigma_x\sigma_y + A^2\sigma_x^2$$

• $e$ is minimum if

$$A = \frac{r\sigma_y}{\sigma_x}; \quad r\sigma_x\sigma_y = C_{xy}$$

• Inserting $A$ into the preceding quadrant (in $e$) results

$$e_m = \sigma_y^2 \left( 1 - r^2 \right)$$
Nonhomogeneous LE

• $AX + B$ is a nonhomogeneous linear estimate of $Y$ in terms of $X$.

• If $Y$ is estimated by a straight line $aX$ passing through the origin, the estimate is homogeneous.

• The random variable $X$ is the data of the estimation; the random variable $\varepsilon = Y - (AX + B)$ is the error of the estimation; and the number $e = E\{\varepsilon^2\}$ is the MS error.
The Orthogonally Principle

• The MS error, $e$, is a function of $A$ and $B$ and it is minimum if $\frac{\partial e}{\partial A} = 0$ and $\frac{\partial e}{\partial B} = 0$.

• The first equation yields

$$\frac{\partial e}{\partial A} = E\{2[Y - (AX + B)](-X)\} = 0$$

$$E\{[Y - (AX + B)](X)\} = 0$$

• This implies that the linear estimate is optimal when the error is orthogonal to the data $X$. 
Homogeneous LE

• We wish to find a constant $a$ such that, if $Y$ is estimated by $aX$, the resulting MS error is minimum

$$Y = aX$$

• We maintain that $a$ must be chosen so that: $E\{ (Y - aX)X \} = 0$

• Since $e = E\{ (Y - aX)^2 \}$, derivation with respect to $a$ and setting the result to 0 yields:

$$\frac{de}{da} = \frac{d}{da} E\{ (Y - aX)(Y - aX) \} = 0$$

$$= \frac{d}{da} E\{ Y^2 + (-2aYX) + a^2(X^2) \} = 0$$

$$E\{ (Y - aX)X \} = 0$$
Homogeneous LE

• The linear estimation of $Y$ in terms of $X$ will be denoted by

$$\hat{E}(Y \mid X) = aX \quad a = \frac{E(XY)}{E(X^2)}$$

• Rewriting the error term:

$$e = E\left\{ (Y - aX)(Y - aX) \right\}$$
$$= E\left\{ (Y^2 + (-2aYX) + a^2(X^2)) \right\}$$
$$= E\left\{ (Y - aX)Y \right\} - E\left\{ (Y - aX)aX \right\}$$
$$= E\left\{ Y^2 \right\} - E\left\{ (aX)^2 \right\} - 2a E\left\{ (Y - aX)X \right\}$$
$$e = E\left\{ (Y - aX)X \right\} = E\left\{ Y^2 \right\} - E\left\{ (aX)^2 \right\}$$
Homogeneous LE

• The last expression is similar to the Pythagorean theorem and is consistent with the orthogonality principle, i.e., the error $Y - aX$ is orthogonal to $X$. 

\[ (Y - aX) \perp X \]
The linear MS estimate of $S$ in terms of the random variables $X_i$ is the sum

$$\hat{S} = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

The constants $a_1, \ldots, a_n$ are so chosen as to make the estimation error $P$ minimum.

$$P = E \left\{ \left( S - \hat{S} \right)^2 \right\} = E \left\{ \left[ S - (a_1 X_1 + \ldots + a_n X_n) \right]^2 \right\}$$

According to the orthogonality principle, $P$ is minimum if the error is orthogonal to the data $X_i$.

$$\frac{\partial P}{\partial a_i} = E \left\{ -2 \left[ S - (a_1 X_1 + \ldots + a_n X_n) \right] X_i \right\} = 0$$

$$E \left\{ \left[ S - (a_1 X_1 + \ldots + a_n X_n) \right] X_i \right\} = 0 \quad \text{for} \quad i = 1, \ldots, n$$
Projection Theorem

- This important result is known also as the projection theorem.
- For $i = 1, \ldots, n$, we obtain:
  \[
  R_{11}a_1 + R_{21}a_2 + \ldots + R_{n1}a_n = R_{01}
  \]
  \[
  R_{12}a_1 + R_{22}a_2 + \ldots + R_{n2}a_n = R_{02}
  \]
  ....................................................
  \[
  R_{1n}a_1 + R_{2n}a_2 + \ldots + R_{nn}a_n = R_{0n}
  \]
  where $R_{ij} = E\{X_iX_j\}$ and $R_{0j} = E\{SX_j\}$
- With $X = [X_1, \ldots, X_n]$, $A = [a_1, \ldots, a_n]$, $R_0 = [R_{01}, \ldots, R_{0n}]$.
  and $R = E\{X^TX\}$, $AR = R_0$ $A = R_0R^{-1}$. 
Projection Theorem

• Inserting the constants \( a_i \) so determined into \( P \), we obtain the Least Mean Square (LMS) error.

• Since \( S - \hat{S} \perp X_i \) for every \( i \), we conclude that

\[
P = E\left\{ (S - \hat{S})S \right\} = E\left\{ S^2 \right\} - AR_0^T
\]