# ECE 2521: Analysis of Stochastic Processes

Lecture 7

Department of Electrical and Computer Engineering
University of Pittsburgh

*October, $27^{th}$ 2021*

Azime Can-Cimino

# One Function of Two Random Variables

Let $W = g(X, Y)$ be a function of RVs $X$ and $Y$

- *Discrete Random Variables*
  - If $X$ and $Y$ are discrete RVs, then $W$ will also be a discrete random variable characterized by a PMF $p_W(w)$
  - The PMF $p_W(w)$ can be obtained by adding the values of $p_{X,Y}(x, y)$ corresponding to $x$ and $y$ pairs for which $g(x, y) = w$: $p_W(w) = \sum_{\{(x,y)|g(x,y)=w\}} p_{X,Y}(x, y)$
- *Continuous Random Variables*
  - If $X$ and $Y$ are continuous RVs and $g(X, Y)$ is a continuous function, then $W = g(X, Y)$ is also a continuous RV
  - To find the PDF $f_W(w)$ of $W$ first find CDF $F_W(w)$ and then take its derivative:
    $F_W(w) = \text{Prob}(W \leq w) = \iint_{g(x,y)\leq w} f_{X,Y}(x, y)dxdy$

# Example

- Let $X$ and $Y$ be any continuous random variables

(1) Determine the PDF of $Z = X + Y$

(2) What if $X$ and $Y$ are independent?

(3) Consider the case when $X$ and $Y$ are independent and uniformly distributed random variables:

$$f_X(x) = u(x) - u(x-1)$$

$$f_Y(y) = 0.5u(y) - 0.5u(y-2)$$

Calculate and plot the PDF of $Z = X + Y$.

# Two Functions of Two Random Variables

- Let $g(X, Y)$ and $h(X, Y)$ be continuous and differentiable functions such that:
$$g(X, Y) = Z \quad \text{and} \quad h(X, Y) = W. \tag{1}$$
- For a given $(z, w)$, (1) may have many solutions. Let $(x_1, y_1), \ldots, (x_n, y_n)$ represent these multiple solutions, such that $g(x_i, y_i) = z$ and $h(x_i, y_i) = w$ for $i = 1, \ldots, n$. Then:
$$f_{ZW}(z, w) = \sum_{i=1}^{n} \frac{1}{|J(x_i, y_i)|} f_{XY}(x_i, y_i)$$
where $x_i = g_I(z, w)$ and $y_i = h_I(z, w)$, and $|J(x_i, y_i)|$ is the determinant of the Jacobian of the transform given in (1) such that:
$$|J(x_i, y_i)| = \begin{vmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{vmatrix}$$

## Example

- Let $Z = aX + bY$ and $W = cX + dY$ are two functions of random variables $X$ and $Y$.
- The joint pdf of $X$ and $Y$ is given by $f_{XY}(x, y)$.
- Find the joint pdf of $Z$ and $W$, $f_{ZW}(z, w)$

# Bivariate Gaussian Random Variables

- Let $X$ and $Y$ be two Gaussian random variables with correlation coefficient $\rho_{XY} = \rho$, where $-1 \le \rho \le 1$
- Their joint probability density function (PDF) is completely characterized by the mean $\mu_X$ and standard deviation $\sigma_X$ of random variable $X$, mean $\sigma_Y$ and standard deviation $\sigma_Y$ of random variable $Y$, and their correlation coefficient $\rho$:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[\frac{\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}}{-2(1-\rho^2)}\right]$$

- If $X$ and $Y$ uncorrelated $\rho = 0$, their joint PDF becomes:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left[-\frac{(x-\mu_X)^2}{2\sigma_X^2} - \frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right] = f_X(x)f_Y(y)$$

- The above implies that uncorrelated Gaussian random variables are also independent.

# Conditional Gaussian PDF

- If $X$ and $Y$ are bivariate Gaussian random variables, the conditional PDF of $X$ given $Y = y$ is:

$$f_{X|Y=y}(x) = \frac{1}{\sigma_X\sqrt{2\pi(1-\rho^2)}}\exp\left[\frac{\left(x-\mu_X-\rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2}{2\sigma_X^2(1-\rho^2)}\right]$$

- The conditional mean of random variable $X$ given $Y = y$ is:

$$E[X|Y=y] = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)$$

- The corresponding conditional variance of $X$ is:

$$\text{Var}(X|Y=y) = \sigma_X^2(1-\rho^2).$$

# Exercise 1

Rectangular to Polar coordinate transformation

- $X, Y \sim \mathcal{N}[0, 1]$ are independent jointly Gaussian random variables
- $R = \sqrt{X^2 + Y^2}$ such that $r = g(x, y) = \sqrt{x^2 + y^2}$
- $\Phi = tan^{-1}\left(\frac{Y}{X}\right)$ such that $\phi = h(x, y) = tan^{-1}\left(\frac{y}{x}\right)$
- Find PDFs of $R$ and $\Phi$.

# Exercise 2

- Let $Z = max(X, Y)$ and $W = min(X, Y)$.
- Determine the PDFs $f_Z(z)$ and $f_W(w)$:

$$
\begin{aligned}
z &= max(x, y) = \begin{cases} x & \text{if } x > y \\ y & \text{if } x \leq y \end{cases} \\
w &= min(x, y) = \begin{cases} y & \text{if } x > y \\ x & \text{if } x \leq y \end{cases}
\end{aligned}
$$

# Probability Models of Multiple Random Variables

- In Chapter 6 we introduce the probability measures for multiple random variables
- A *vector random variable* X is a function that assigns a vector of real numbers to each outcome $\xi$ in $S$, the sample space of the random experiment:

$$\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T : S \to \mathbb{R}^n$$

- The probability models of $n$ random variables are the generalization of the probability models of two random variables.

# Probability Models of Multiple Random Variables

- A **random vector** is a column vector $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T$, where each $X_i$ is a random variable: when $n = 1$ a random vector reduces to a random variable
- A **sample value of a random vector** is a column vector $\mathrm{x} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^T$, where each $x_i$ is a sample value of the random variable $X_i$
- **Random vector probability functions**:
  (a) The CDF of a random vector X is
  $$F_{\mathrm{X}}(\mathrm{x}) = F_{X_1,\dots,X_n}(x_1, \dots, x_n)$$
  (b) The PMF of a discrete random vector X is
  $$p_{\mathrm{X}}(\mathrm{x}) = p_{X_1,\dots,X_n}(x_1, \dots, x_n)$$
  (c) The PDF of a continuous random vector X is
  $$f_{\mathrm{X}}(\mathrm{x}) = f_{X_1,\dots,X_n}(x_1, \dots, x_n)$$

# Multivariate Joint CDF

- The joint CDF of random variables $X_1, \ldots, X_n$ is

$$F_{\mathrm{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

- The joint CDF is defined for discrete, continuous, and mixed type random variables
- **Properties**
  (1) $0 \leq F_{\mathrm{X}}(\mathrm{x}) \leq 1$.
  (2) $F_{\mathrm{X}}(x_1, \ldots, x_n)$ is nondecreasing on all $x_i$ for $i = 1, \ldots, n$.
  (3) $\lim_{x_1 \to -\infty, \ldots, x_n \to -\infty} F_{\mathrm{X}}(x_1, \ldots, x_n) = 0$.
  (4) $\lim_{x_1 \to \infty, \ldots, x_n \to \infty} F_{\mathrm{X}}(x_1, \ldots, x_n) = 1$.
  (5) Joint CDF for $X_1, \ldots, X_{n-1}$ is given by $F_{X_1, \ldots, X_n}(x_1, \ldots, x_{n-1}, \infty)$.

# Multivariate Joint PMF

- The joint PMF of discrete random variables $X_1, \ldots, X_n$:

$$p_{\mathrm{X}}(\mathrm{x}) = p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathrm{Prob}\left[X_1 = x_1, \ldots, X_n = x_n\right]$$

- Satisfies the axioms of probability:
  (a) Non-negativity: $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \geq 0$
  (b) Normalization: $\sum_{x_1} \cdots \sum_{x_n} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = 1$

- Probability of an event $A$ is given by:

$$\mathrm{P}\left[A\right] = \sum_{(x_1,\ldots,x_n)\in A} \cdots \sum p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \qquad X_1, \ldots, X_n \text{ discrete}$$

# Multivariate Joint PMF

- Marginal PMFs:

$$\begin{aligned} p_{X_1,\ldots,X_{n-1}}(x_1,\ldots,x_{n-1}) &= \sum_{x_n} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \\ p_{X_1}(x_1) &= \sum_{x_2}\cdots\sum_{x_n} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \end{aligned}$$

- Conditional PMFs:

$$p_{X_n}(x_n|x_1,\ \ldots,\ x_{n-1}) = \frac{p_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{p_{X_1,\ldots,X_{n-1}}(x_1,\ldots,x_{n-1})}$$

- Recursively, we can obtain:

$$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = p_{X_n}(x_n|x_1,\ \ldots,\ x_{n-1})$$
$$p_{X_{n-1}}(x_{n-1}|x_1,\ \ldots,\ x_{n-2})\cdots p_{X_2}(x_2|x_1)p_{X_1}(x_1)$$

# Multivariate Joint PDF

- The joint PDF of continuous random variables $X_1, \ldots, X_n$ is denoted by $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$, where:

$$\text{Prob}\left[a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n\right] =$$

$$\int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

- Satisfies the axioms of probability:
  (a) Non-negativity: $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \geq 0$
  (b) Normalization: $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n = 1$
- Probability of an event $A$ is given by:

$$\mathrm{P}\left[A\right] = \underset{(x_1,\ldots,x_n)\in A}{\int \cdots \int} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n \qquad X_1, \ldots, X_n \text{ continuous}$$

# Multivariate Joint PDF

- Marginal PDFs:

$$f_{X_1,\ldots,X_{n-1}}(x_1,\ldots,x_{n-1}) = \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\ dx_n$$

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\ dx_2 \ldots dx_n$$

- Conditional PDFs

$$f_{X_n}(x_n|x_1,\ \ldots,\ x_{n-1}) = \frac{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{f_{X_1,\ldots,X_{n-1}}(x_1,\ldots,x_{n-1})}$$

# Multivariate Joint PDF

- Then recursively, we can obtain:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_n}(x_n|x_1,\ \ldots,\ x_{n-1})\cdot f_{X_{n-1}}(x_{n-1}|x_1,\ \ldots,\ x_{n-2})\cdots f_{X_2}(x_2|x_1)f_{X_1}(x_1)$$

$f_{x_3}(x_3|x_1,x_2)\ \ldots$

Note:

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \int_{-\infty}^{x_n}\cdots\int_{-\infty}^{x_1} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)dx_1\ldots dx_n$$

- Therefore:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{\partial^n}{\partial x_1\ldots\partial x_n}F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$$

# Example

- Random variables $X_1, \ldots, X_n$ have joint PDF:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \begin{cases} 1 & 0 \le x_i \le 1, \; i = 1,\ldots,n \\ 0 & \text{otherwise} \end{cases}$$

- Let $A$ denote the event that $\max_i X_i \le \frac{1}{2}$
- Find $P[A]$. All should be less than $\frac{1}{2}$

$P[x_1 \le \frac{1}{2}, \ldots, x_n \le \frac{1}{2}]$

$= \int_0^{\frac{1}{2}} \cdots \int_0^{\frac{1}{2}} 1 \, dx_1 \ldots dx_n = \frac{1}{2^n}$

# Example

- Random variables $X_1, \dots, X_n$ have joint PDF:

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \begin{cases} 1 & 0 \le x_i \le 1,\ i = 1,\dots,n \\ 0 & \text{otherwise} \end{cases}$$

- Let $A$ denote the event that $\max_i X_i \le \frac{1}{2}$
- Find $P[A]$.

Solution The maximum of $n$ numbers is less than $\frac{1}{2}$ if and only if each of the $n$ numbers is less than $\frac{1}{2}$; therefore

$$\begin{aligned} P[A] &= P\left[\max_i X_i \le \frac{1}{2}\right] = P\left[X_1 \le \frac{1}{2}, \dots, X_n \le \frac{1}{2}\right] \\ &= \int_0^{\frac{1}{2}} \dots \int_0^{\frac{1}{2}} 1\ dx_1 \dots dx_n = \frac{1}{2^n} \end{aligned}$$

# Independence

Sub-index assumes different distributions

- $X_1, \ldots, X_n$ are **independent** if for all $x_1, \ldots, x_n$:

Becomes multiplication

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\ldots p_{X_n}(x_n)$$
$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$$

- $X_1, \ldots, X_n$ are **Independent Identically Distributed (i.i.d)** if for all $x_1, \ldots, x_n$:

Have same distribution

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_X(x_1)p_X(x_2)\ldots p_X(x_n)$$
$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_X(x_1)f_X(x_2)\ldots f_X(x_n)$$

# Example

- The random variables $X_1$,$X_2$ and $X_3$ have the joint Gaussian PDF:

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = \frac{e^{-(x_1^2+x_2^2-\sqrt{2}x_1x_2+\frac{1}{2}x_3^2)}}{2\pi\sqrt{\pi}}$$

- Find the marginal PDFs $f_{X_1,X_3}(x_1, x_3)$, $f_{X_1}(x_1)$ and $f_{X_3}(x_3)$.

# Example - Solution

- The marginal PDF for the pair $X_1$ and $X_3$ is found by integrating the joint PDF over $X_2$:

$$f_{X_1,X_3}(x_1,x_3) = \frac{e^{-x_3^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-(x_2^2 - \sqrt{2}x_1x_2 + \frac{x_1^2}{2} + \frac{x_1^2}{2})}}{\pi\sqrt{2}} dx_2 = \frac{e^{-x_3^2/2}}{\sqrt{2\pi}} \frac{e^{-x_1^2/2}}{\sqrt{2\pi}}$$

- Marginal PDF for $X_1$ is found by integrating $f_{X_1,X_3}(x_1,x_3)$ over $X_3$:

All Gaussian → marginals are Gaussian

$$f_{X_1}(x_1) = \frac{e^{-x_1^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-x_3^2/2}}{\sqrt{2\pi}} dx_3 = \frac{e^{-x_1^2/2}}{\sqrt{2\pi}}$$

- Marginal PDF for $X_3$ is found by integrating $f_{X_1,X_3}(x_1,x_3)$ over $X_1$:

$$f_{X_3}(x_3) = \frac{e^{-x_3^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-x_1^2/2}}{\sqrt{2\pi}} dx_1 = \frac{e^{-x_3^2/2}}{\sqrt{2\pi}}$$

Note: $f_{X_1,X_3}(x_1,x_3) = f_{X_1}(x_1) f_{X_3}(x_3)$, therefore $X_1$ and $X_3$ are independent zero-mean, unit variance Gaussian random variables.

# Functions of Random Vectors

- Let $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T$ and $Y = g(\mathrm{X})$; that is $g : \mathbb{R}^n \to \mathbb{R}$ with $\mathrm{X} \in \mathbb{R}^n$ and $Y \in \mathbb{R}$. Then:

$$F_Y(y) = Prob(g(\mathrm{X}) \leq y) = Prob(X \in R_Y)$$

where $R_Y = \{\mathrm{x} : g(\mathrm{x}) \leq y\}$.

# Transformations of Random Vectors

- Consider the random vector: $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T$
- Let $\mathrm{Y} = \mathrm{g}(\mathrm{X}) = [g_1(\mathrm{X}) \cdots g_n(\mathrm{X})]^T$ such that $\mathrm{g} : \mathbb{R}^n \to \mathbb{R}^n$ with $\mathrm{X} \in \mathbb{R}^n$ and $\mathrm{Y} \in \mathbb{R}^n$
- If $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T = \mathrm{g}^{-1}(\mathrm{Y}) = [g_1^{-1}(\mathrm{Y}) \cdots g_n^{1}(\mathrm{Y})]^T$, we can compute $f_{\mathrm{Y}}(\mathrm{y})$ as:

$$f_{\mathrm{Y}}(\mathrm{y}) = \frac{f_{\mathrm{X}}(\mathrm{g}^{-1}(\mathrm{y}))}{|J(x_1, \dots, x_n)|}$$

where $|J(x_1, \dots, x_n)|$ is the determinant of the Jacobian:

$$J(x_1, \dots, x_n) = \begin{bmatrix} \frac{\partial g_1(\mathrm{x})}{\partial x_1} & \dots & \frac{\partial g_1(\mathrm{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_n(\mathrm{x})}{\partial x_1} & \dots & \frac{\partial g_n(\mathrm{x})}{\partial x_n} \end{bmatrix}$$

# Special Case (Linear Transformation)

- Let $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T$ and $\mathrm{Y} = \mathrm{g}(\mathrm{X}) = \boldsymbol{A}\mathrm{X} + \mathrm{b}$ where $\boldsymbol{A}$ is an invertible $n \times n$ matrix and b is an $n \times 1$ vector
- Then $\mathrm{X} = \boldsymbol{A}^{-1}(\mathrm{Y} - \mathrm{b})$ and:

$$f_{\mathrm{Y}}(\mathrm{y}) = \frac{f_{\mathrm{X}}(\boldsymbol{A}^{-1}(\mathrm{Y} - \mathrm{b}))}{|\boldsymbol{A}|}$$

# Expected Values of Random Vectors

- Let $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T$ and $Y = g(\mathrm{X}) = g(X_1, \dots, X_n)$, then the expected value of $Y$ is:

$$E[Y] = \begin{cases} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathrm{X}) f_{\mathrm{X}}(x_1, \dots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n & \mathrm{X} \text{ is jointly continuous} \\ \sum_{x_1} \cdots \sum_{x_n} g(\mathrm{X}) p_{\mathrm{X}}(x_1, \dots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n & \mathrm{X} \text{ is jointly discrete} \end{cases}$$

# Mean Vector

- Let $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T$, then the expected value of X - also the mean vector $\mu_\mathrm{X}$ - is defined as:

$$\mu_\mathrm{X} = E[\mathrm{X}] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$$

- In general if $\mathrm{Y} = \mathrm{g}(\mathrm{X}) = [g_1(\mathrm{X}) \cdots g_n(\mathrm{X})]^T$, then the expected value of Yis computed as:

$y = ax + b$
$E[y] = a E[x] + b$

$$E[\mathrm{g}(\mathrm{X})] = \begin{bmatrix} E[g_1(\mathrm{X})] \\ \vdots \\ E[g_n(\mathrm{X})] \end{bmatrix}$$

# Covariance Matrix and Correlation Matrix

- The correlation matrix $\boldsymbol{R}_{\mathrm{X}} = E[\mathrm{X}\mathrm{X}^T]$:

$$\boldsymbol{R}_{\mathrm{X}} = \begin{bmatrix} E[X_1^2] & \cdots & E[X_1 X_n] \\ \vdots & & \vdots \\ E[X_n X_1] & \cdots & E[X_n^2] \end{bmatrix}$$

- The covariance matrix $\boldsymbol{K}_{\mathrm{X}} = E[(\mathrm{X} - \mu_{\mathrm{X}})(\mathrm{X} - \mu_{\mathrm{X}})^T]$:

$$\boldsymbol{K}_{\mathrm{X}} = \begin{bmatrix} E[(X_1 - \mu_1)^2 & E[(X_1 - \mu_1)(X_2 - \mu_2) & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1) & E[(X_2 - \mu_2)^2 & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & E[(X_n - \mu_n)^2] \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n)] \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix}$$

Note: $\boldsymbol{K}_{\mathrm{X}} = \boldsymbol{R}_{\mathrm{X}} - \boldsymbol{\mu}_{\mathrm{X}} \boldsymbol{\mu}_{\mathrm{X}}^T$.

# Theorem

- For a linear transformation of a vector of random variables of the form $\mathrm{Y} = \boldsymbol{A}\mathrm{X} + \mathrm{b}$, the means of X and **Y** are related by:

$$\mu_{\mathrm{Y}} = \boldsymbol{A}\mu_{\mathrm{X}} + \mathrm{b}$$

- Also, the covariance matrices of X and Y are related by:

$$\boldsymbol{K}_{\mathrm{Y}} = \boldsymbol{A}\boldsymbol{K}_{\mathrm{X}}\boldsymbol{A}^{T}.$$

# Remarks

- Both $\boldsymbol{R}_{\mathrm{X}}$ and $\boldsymbol{K}_{\mathrm{X}}$ are symmetric nonnegative definite $n \times n$ matrices.
- Recall from linear algebra that, if $\mathrm{u_i}$ for $i = 1, \ldots, n$ are eigenvectors with the corresponding eigenvalues $\lambda_i$ with $\lambda_i \geq 0$ such that $\boldsymbol{K}_{\mathrm{X}}\mathrm{u}_i = \lambda_i \mathrm{u}_i$ and $\mathrm{u}_i$'s are orthogonal, then:

SVD | Eigenvalue Decomposition

$$\boldsymbol{K}_{\mathrm{X}} = \boldsymbol{U}\Lambda\boldsymbol{U}^T$$

  where $\boldsymbol{U} = [\mathrm{u}_1 \cdots \mathrm{u}_n]$ is an orthogonal matrix with $i^{th}$ eigenvector as the $i^{th}$ column ($\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$), and $\Lambda$ is a diagonal matrix with $i^{th}$ diagonal element as the $i^{th}$ eigenvalue $\lambda_i$.
- Given $\mathrm{Y} = \boldsymbol{A}\mathrm{X}$, we can choose $\boldsymbol{A}$ such that Y has uncorrelated components: $\boldsymbol{A} = (\boldsymbol{U}\sqrt{\Lambda})^{-1}$ yields $\boldsymbol{K}_{\mathrm{Y}} = \boldsymbol{I}$.

# Joint Moment Generating Functions of Random Vectors

- Let $\mathrm{X} = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}^T$, then the joint moment generating function of X is defined as:

$$\Phi_{\mathrm{X}}(\mathrm{s}) = \Phi_{X_1,\ldots,X_n}(s_1, \ldots, s_n) = E[e^{\mathrm{s}^T\mathrm{X}}] = E[e^{s_1X_1+\ldots+s_nX_n}]$$

where $\mathrm{s} = [s_1 \ldots s_n]^T$.

- The joint PDF can be obtained using the MGF of X:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^n}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Phi_{\mathrm{X}}(\mathrm{s})e^{s_1X_1+,\ldots,s_nX_n}ds_1\cdots ds_n$$

- Recall that for $\mathrm{s} = j\omega = [j\omega_1 \ldots j\omega_n]$ we can compute the joint characteristic function of X
- If $X_1, \ldots, X_n$ are all independent, then:

$$\Phi_{\mathrm{X}}(\mathrm{s}) = \Phi_{X_1}(s_1)\cdots\Phi_{X_n}(s_n) = \prod_{i=1}^{n}\Phi_{X_i}(s_i)$$

# Multivariate Gaussian Random Variables

- If a random vector $\mathrm{X} = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}^T \in \mathbb{R}^n$ is said to follow a multivariate Gaussian distribution with mean $\mu_{\mathrm{X}}$ and covariance $\boldsymbol{K}_{\mathrm{X}}$ (where $\boldsymbol{K}_{\mathrm{X}}$ is invertible), then

$$f_{\mathrm{X}}(\mathrm{X}) = (2\pi)^{-\frac{n}{2}} (\det \boldsymbol{K}_{\mathrm{X}})^{-\frac{1}{2}} \exp\left[-\frac{(\mathrm{X}-\mu_{\mathrm{X}})^T \boldsymbol{K}_{\mathrm{X}}^{-1} (\mathrm{X}-\mu_{\mathrm{X}})}{2}\right]$$

## Properties

(1) Uncorrelated Gaussian random variables are independent. That is, if $X$ and $Y$ are jointly Gaussian and $E[(X - \mu_X)(Y - \mu_Y)] = 0$, then $X$ and $Y$ are independent.

(2) If $\mathrm{X} \in \mathbb{R}^n$ follows a multivariate Gaussian distribution, then $\mathrm{Y} = \boldsymbol{A}\mathrm{X} + \mathrm{b}$ with $\boldsymbol{A}$ as an $n \times n$ matrix and b as an $n \times 1$ vector also follows a multivariate Gaussian distribution. That is $\mathrm{Y} \sim \mathcal{N}(\boldsymbol{A}\mathrm{X} + \mathrm{b}, \boldsymbol{A}^T \boldsymbol{K}_\mathrm{X} \boldsymbol{A})$ $K_x = I$ for iid

(3) All the marginal distributions are also Gaussian. That is, $X_i$ for $i = 1, \ldots, n$ also follows a Gaussian distribution. That is $X_i \sim \mathcal{N}\left(\mu_i, \mathrm{Var}(X_i)\right)$.

# Properties

(4) If we denote $\mathrm{X} = \begin{bmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \end{bmatrix}$ such that $\mu_{\mathrm{X}} = \begin{bmatrix} \mu_{\mathrm{X}_1} \\ \mu_{\mathrm{X}_2} \end{bmatrix}$ and $\boldsymbol{K}_{\mathrm{X}} = \begin{bmatrix} \boldsymbol{K}_{\mathrm{X}_1} & \boldsymbol{K}_{\mathrm{X}_1,\mathrm{X}_2} \\ \boldsymbol{K}_{\mathrm{X}_2,\mathrm{X}_1} & \boldsymbol{K}_{\mathrm{X}_2} \end{bmatrix}$, then the conditional random variable $\mathrm{X}_1|\mathrm{X}_2$ also follows a Gaussian distribution such that $\mathrm{X}_1|\mathrm{X}_2 \sim \mathcal{N}\left(\mu_{\mathrm{X}_1} + \boldsymbol{K}_{\mathrm{X}_1,\mathrm{X}_2}\boldsymbol{K}_{\mathrm{X}_2}^{-1}(\mathrm{x}_2 - \mu_{\mathrm{X}_2}), \boldsymbol{K}_{\mathrm{X}_1} - \boldsymbol{K}_{\mathrm{X}_1,\mathrm{X}_2}\boldsymbol{K}_{\mathrm{X}_2}^{-1}\boldsymbol{K}_{\mathrm{X}_2,\mathrm{X}_1}\right)$.

(5) The joint MGF of X: $\Phi_{\mathrm{X}} = \exp\left(\mathrm{s}^T\mu_{\mathrm{X}} + \frac{1}{2}\mathrm{s}^T\boldsymbol{K}_{\mathrm{X}}\mathrm{s}\right)$.

## Estimation versus Detection

- Main difference between estimation and detection problems involves how we measure success:

**Detection** We might ask how often our guess is correct

**Estimation** Common to measure an error between the true value and the estimated value.

- In detection problems, we are interesting in estimating a quantity that is discrete in nature:

**Example 1** Radar systems: we are trying to decide whether or not a target is present based on observing radar returns

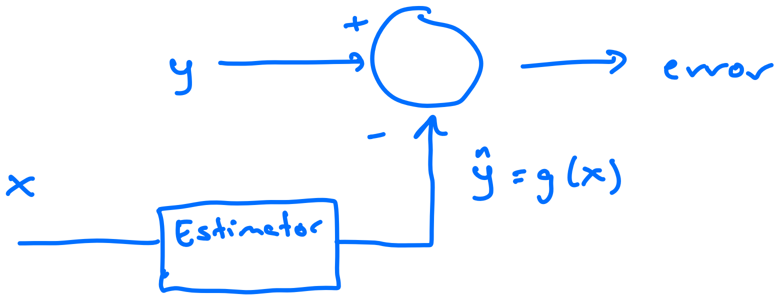
**Example 2** Digital communication systems: we are trying to determine whether bits take on values of 0 or 1 based on samples of some receive signal

# Estimation

If $x, y$ are correlated to some degree

$x$ is observed, estimate $y$ MSE (mean-square error)

$$E[y - g(x)]^2$$




## Mean-square estimation

Estimator $\hat{Y} = g(x)$ : find $g(\ )$ that minimizes $E[Y - g(\overset{\hat{Y}}{x})]^2$

$$\boxed{\text{Solution: } \hat{y} = E[y \mid x]}$$

$$e = E[y - g(x)]^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - g(x))^2 \, \underbrace{f_{xy}(x,y)}_{f_x(x) \cdot f_y(y \mid x)} \, dx \, dy$$

$$= \int_{-\infty}^{\infty} f_x(x) \left[ \int_{-\infty}^{\infty} (y - g(x))^2 \, f_y(y \mid x) \, dy \right] dx$$

Really want to minimize

$$\int_{-\infty}^{\infty} (y - g(x))^2 \, f_y(y \mid x) \, dy \longrightarrow \text{Take derivative w.r.t. } g(x)$$

$$-2 \int_{-\infty}^{\infty} (y - g(x)) \, f_y(y \mid x) \, dy = 0$$

$$g(x) \underbrace{\int_{-\infty}^{\infty} f_y(y|x)\, dy}_{=1 \text{ since this is a PDF}} = \int_{-\infty}^{\infty} y \, f_y(y|x)\, dy$$

$$\therefore \boxed{g(x) = E[y|x]}$$

**Remarks**

> $E[Y|X]$ is (in general) nonlinear function of $X$, nonlinear estimator

> If $X$ and $Y$ are Gaussian, $E[Y|X]$ is a linear function of $X$

> If $y = h(x) \Rightarrow E(y = H(x)|x) = h(x)$

> If $x, y$ are independent $\Rightarrow E[y|x] = E[y]$

> $e_{min} = \boxed{E[y - E[y|x]]^2}$ : conditional variance of $y$ given $x$

> $g^*(x) = E[y|x]$ is the best approximation of $y$ in mean square sense

$$E(y - g(x))^2 \geq E(y - E[y|x])^2 \quad \text{for all } g(\cdot)$$

MMSE : minimal mean square estimator

$$g(x) = \hat{y} = E[y|x]$$

**Proof:** $E[y - g(x)]^2 = E[y - E[y|x] + E[y|x] - g(x)]^2$

$$= E[y - \underbrace{E(y|x)}_{g^*(x)}]^2 + E[E(y|x) - g(x)]^2$$

$$+ 2E[(y - E(y|x))(E[y|x] - g(x))]$$

show that this is 0

$$h(x) = E[y|x] - g(x)$$

$$E[(y - E[y|x])h(x)] = E[y\,h(x)] - E[h(x)\,E[y|x]]$$

$$E[h(x)\,y] = E[h(x)\,E[y|x]]$$

## Iterated expectation

$E[y|X]$ is just function of $x$, $G(x)$

$z = G(x) = E[y|X]$ if you don't specify $x$

$$E[G(x)] = \int_{-\infty}^{\infty} G(x)\, f_x(x)\, dx$$

$$= \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} y \cdot \underline{f_{y|x}(y|x)}\, dy \right] \underline{f_x(x)}\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \cdot f_{xy}(x,y)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} y \int^{\infty} f_{xy}(x,y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} y\, f_x(y)\, dy$$

$$\boxed{E[E[y|x]] = E[y]}$$

$$E[g(x,y)] = E[E[g(x,y) \mid X = x]]$$

$$g(x,y) = h(x)\,y \longrightarrow E[h(x)\,y] = E[E[h(x)\,y \mid x]]$$

$$= E[h(x)\, E[y|x]]$$

## Linear mean-square estimation

$$\hat{y} = \alpha + \beta x$$

$\alpha = 0$ : homogeneous linear

$\alpha \neq 0$ : nonhomogeneous linear

$$\min_{\alpha, \beta} \text{ w.r.t. } \left[ \varepsilon = E[y - \hat{y}]^2 \right]$$

(i) $\dfrac{d\varepsilon}{d\alpha} = 0 = 2E[y - \alpha - \beta x](-1) \longrightarrow E[y] = \alpha + \beta E[x]$

(ii) $\dfrac{d\varepsilon}{d\beta} = 0 = 2E[(y - \alpha - \beta x)(-x)] \longrightarrow E[yx] = \alpha E[x] + \beta E[x^2]$

$$\begin{bmatrix} 1 & E[x] \\ E[x] & E[x^2] \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} E[y] \\ E[yx] \end{bmatrix}$$

$$\beta = \frac{\operatorname{cov}(x, y)}{\underbrace{\operatorname{var}(x)}_{\sigma_x^2}} \qquad \alpha = E[y] - \beta E[x]$$

$$\boxed{\hat{y} = E[y] + [x - E[x]] \frac{r_{xy}\, \sigma_y}{\sigma_x}}$$

LMSE

$$\varepsilon = \sigma_y^2 (1 - r_{xy}^2)$$

Remarks

> Assume $x$ is not measured $\quad \hat{y} = \alpha = E[y]$

$$\varepsilon_{min} = \sigma_y^2$$

$$\min_{\alpha} \varepsilon = E[y - \alpha]^2$$

$$\frac{d\varepsilon}{d\alpha} = 2E[y - \alpha](-1) = 0 \longrightarrow \alpha = E[y]$$

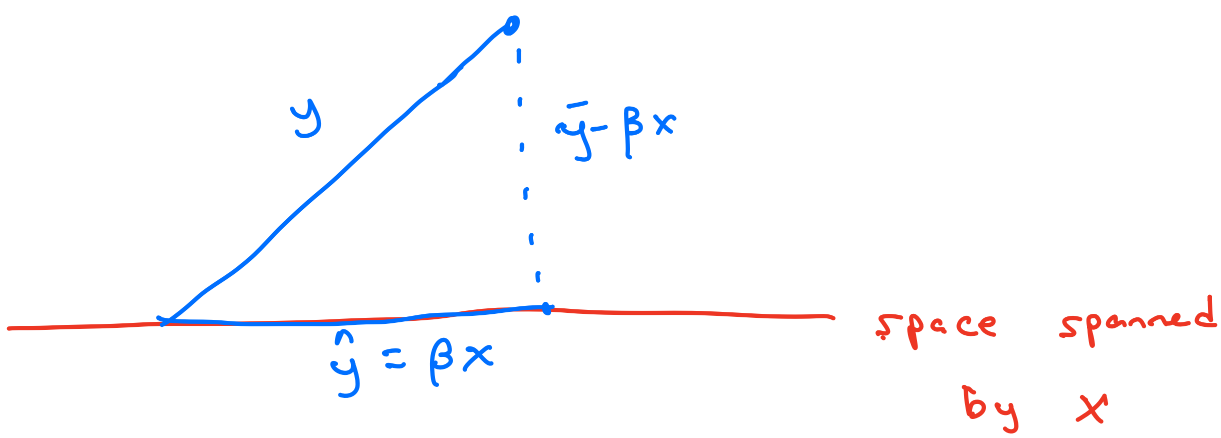
> $x, y$ are uncorrelated

$$\hat{y} = E[y] \qquad \varepsilon_{min} = \sigma_y^2$$

Orthogonality principle:

$$E[\underbrace{(y - \beta x)}_{\text{error}} \cdot \underbrace{x}_{\text{measurements}}] = 0$$

(Zero mean case) $\alpha = 0$



# Maximum A-Posteriori (MAP) Estimator

- Assume $X$ and $Y$ are correlated to some degree
- Find the most probable input $X$ given the observation $Y = y$

Discrete Find the value of $x$ that maximizes the a posteriori probability $P[X = x|Y = y]$:

$$\hat{X}_{MAP} = \max_x P[X = x|Y = y]$$

Cont. $\hat{X}_{MAP} = \max_x f_{X|Y}(x|y)$

Particle filter

knowing y

# Maximum Likelihood (ML) Estimator

**Discrete** The a posteriori probability is given by:

$$P[X = x|Y = y] = \frac{P[Y = y|X = x]P[X = x]}{P[Y = y]}$$

- $P[Y = y]$ does not affect the optimization (ignore)
- The a priori probability $P[X = x]$ may not be known, and we can model it as a uniform distribution (constant)
- Select the estimator $\hat{X}_{ML}$ that maximizes $P[Y = y|X = x]$ as the maximum likelihood (ML) estimator of the observed value $Y = y$:

$$\hat{X}_{ML} = \max_x P[Y = y|X = x]$$

**Cont.** Similarly:

$$\hat{X}_{ML} = \max_x f_{Y|X}(Y|X)$$

# Example

- Find the MAP and ML estimators of $X$ in terms of the observations $Y$ when $X$ and $Y$ are jointly Gaussian random variables with the following conditional PDFs:

MAP

$$f_{X|Y} = \frac{e^{-\frac{1}{2(1-\rho^2)\sigma_X^2}\left(x-\rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)-\mu_X\right)^2}}{\sqrt{2\pi\sigma_X^2(1-\rho^2)}}$$

MLE

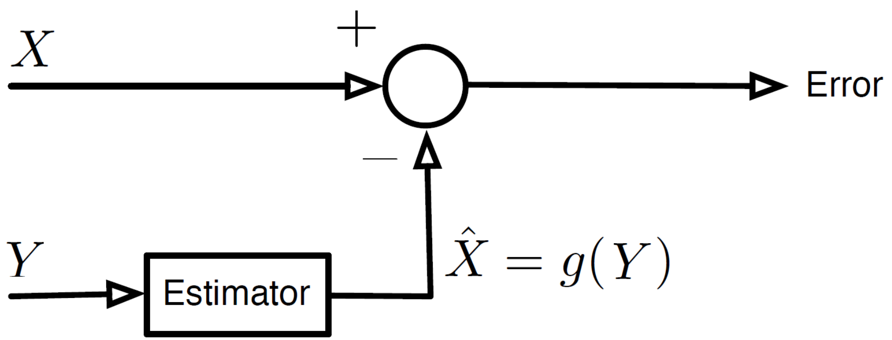
$$f_{Y|X} = \frac{e^{-\frac{1}{2(1-\rho^2)\sigma_Y^2}\left(y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)-\mu_Y\right)^2}}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}}$$

# Estimation of Random Variables

- Estimating the parameters of one or more random variables (e.g. probabilities, means, variances, or covariances)
- Estimating the value of an inaccessible random variable $X$ in terms of the observation of an accessible random variable $Y$:
    - Prediction Problems: predict future based on current and past observations
    - Interpolation Problems: given samples of a signal, we wish to interpolate to some in-between point in time

      Extrapolation is opposite
    - Filtering Problems: filter the noise out of a sequence of observations to provide the best estimate of the desired signal

# Mean-Square Estimation (MSE)




- Assume $X$ and $Y$ are correlated to some degree
- If $Y$ is observed, then estimate $X$ so as to minimize the mean-square error:

$$e = E[(X - g(Y))^2]$$

# Constant MSE

(a) Estimate the random variable $X$ by a constant $\hat{X} = g(Y) = a$ so that the mean-square error is minimized.

(b) What is the mean-square error for this estimator?

# Linear MSE

- Estimate $X$ by a linear function $g(Y) = aY + b$ so that the mean-square error is minimized:

$$\min_{a,b} E[(X - aY - b)^2]$$

Step 1 We can apply the result from the previous example if we view the problem as estimating the random variable $(X - aY)$ with a constant $b$, such that:

$$b^* = E[X - aY] = E[X] - aE[Y]$$

Step 2 The minimization problem simplifies to one parameter $a$:

$$\min_{a} E[(X - E[X] - a(Y - E[Y]))^2]$$

such that $a^* = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}$

# Linear MSE

- The linear estimate $g(Y) = aY + b$ of $X$ is obtained:

$$\hat{X} = E[X] + \mathrm{Cov}(X, Y)\frac{Y - E[Y]}{\mathrm{Var}(Y)}$$

Note The linear mean-square estimator depends on second order moments: mean, variance and covariance.

- The minimum error of the linear MSE:
$\epsilon_{MIN} = \mathrm{Var}(X)\left(1 - \rho^2\right)$.

## Linear Mean Square Estimation

Given: $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho_{XY}$ and an observation of $X$.

Goal: Get an estimate of $Y$ in the form:

$\hat{Y}_{LNH} = aX + b$ Linear non-homogenous (LNH)

$\hat{Y}_{LH} = aX$ Linear homogenous (LH)

Intuition: If $X$ and $Y$ are well correlated, $\hat{Y}_{LNH}$ should be a "good" estimator.

Linear correlation

## Applications

- One step predictor: $x_1, x_2, x_3, \ldots$ is a sequence of correlated random variables (Dow Jones Industrial Average?)

$$\hat{X}_{n+1} = aX_n + b$$

- Weight, $W$, and cholesterol level, $C$

$$\hat{C} = aW + b$$

## Mean Squared Error

Goodness is measured in mean squared error (MSE). Let $\varepsilon$ be the estimation error. Then,

$$MSE = E[\varepsilon^2] = E\left[\left(Y - \hat{Y}\right)^2\right]$$

= "average error power"

Pick coefficients $a$ and $b$ (or just $a$ for homogenous case) to minimize MSE.

## Linear Non-Homogenous Estimation

$$MSE = E\left\{[Y - (aX + b)]^2\right\}$$

$$= E[Y^2] - 2aE[XY] - 2bE[Y] + a^2E[X^2] + 2abE[X] + b^2$$

$$\frac{\partial MSE}{\partial a} = -2E[XY] + 2aE[X^2] + 2bE[X] = 0$$

$$\frac{\partial MSE}{\partial b} = -2E[Y] + 2aE[X] + 2b = 0$$

$$\boxed{a = \frac{\sigma_Y}{\sigma_X}\rho_{XY} \qquad b = E[Y] - aE[X]}$$

## Linear Non-Homogenous Estimation, Cont'd

$$\therefore \hat{Y}_{LNH} = \frac{\sigma_Y}{\sigma_X}\rho_{XY} X + m_Y - am_X$$

Rearrangement yields

$$\hat{Y}_{LNH} = \sigma_Y \rho_{XY} \overbrace{\underbrace{\left(\frac{X - m_X}{\sigma_X}\right)}_{\text{Zero-mean, unit variance version of } X}}^{\text{Zero-mean, unscaled version of } \hat{Y}} + m_Y$$

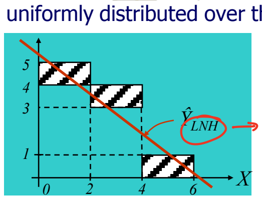
scaling slope offset

normalized random variable

Scaling

## Example of LNH Estimator

Let $X$ and $Y$ be uniformly distributed over the shaded region:




y | X

has offset

Needed moments: $m_X = 3, \sigma_X = \sqrt{3},$

$$m_Y = \frac{17}{6}, \sigma_Y = 1.724, \rho_{XY} = -0.893$$

$$\hat{Y}_{LNH} = -0.889X + 5.5$$

## Orthogonality Condition

Recall the optimal *"a"* for $\hat{Y}_{LNH}$ solves: $\frac{\partial}{\partial a} E[\varepsilon^2] = 0$

$$\frac{d}{da} E[\varepsilon^2] = E\left[2\varepsilon\left(\frac{d}{da}\varepsilon\right)\right]$$
$$= 2E\left\{\varepsilon\left(\frac{d}{da}[Y - a(X - m_X) - m_Y]\right)\right\}$$
$$= 2E\{\varepsilon(X - m_X)\}$$

$$\Rightarrow E[\varepsilon(X - m_X)] = 0$$

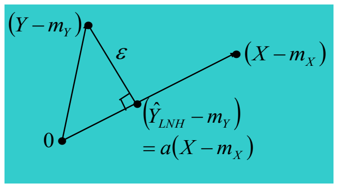
Also, because $E[\varepsilon] = 0$, $E[\varepsilon X] = 0$

*Orthogonality between error and "data"*

$E[(y-\hat{y})(x-\mu_x)] = 0$

## Geometrical View – Non-homogeneous Case

$$(\hat{Y}_{LNH} - m_Y) = a(X - m_X)$$
$$\varepsilon = (Y - m_Y) - a(X - m_X)$$




The estimator is the point in the space spanned by $(X\text{-}m_X)$ that is **nearest** to $(Y\text{-}m_Y)$.

## Orthogonality Condition for the Homogeneous Case

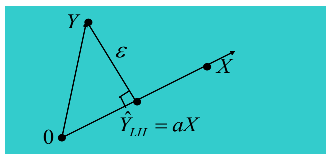
Recall the optimal *"a"* for $\hat{Y}_{LH}$ solves: $\frac{\partial}{\partial a} E[\varepsilon^2] = 0$

$$\frac{d}{da} E[\varepsilon^2] = E\left[2\varepsilon\left(\frac{d}{da}\varepsilon\right)\right]$$
$$= 2E\left\{\varepsilon\left(\frac{d}{da}[Y - aX]\right)\right\}$$
$$= 2E\{\varepsilon(X)\}$$

*Same expression as on a previous slide, but here, derived directly*

## Geometrical View – Homogeneous Case

$$\hat{Y}_{LH} = aX$$
$$\varepsilon = Y - aX$$




The estimator is the point in the space spanned by $X$ that is **nearest** to $Y$.

## Performance of $\hat{Y}_{LNH}$

$$MSE_{opt} = E\{\varepsilon[(Y - m_Y) - a(X - m_X)]\}$$

*orthogonal*

$$= E\{\varepsilon(Y - m_Y)\}$$
$$= E\{[(Y - m_Y) - a(X - m_X)](Y - m_Y)\}$$
$$= \sigma_Y^2 - a\,\mathrm{cov}(X,Y)$$
$$= \sigma_Y^2 - \frac{\sigma_Y}{\sigma_X}\rho_{XY}\,\mathrm{cov}(X,Y)$$
$$= \sigma_Y^2(1 - \rho_{XY}^2)$$

Lowest error

## Observations About Optimal MSE

$$MSE_{opt} = \sigma_Y^2(1 - \rho_{XY}^2)$$

- Lowest when $|\rho_{XY}| = 1$ (Perfect correlation$\Rightarrow Y = aX + b$)
- Highest when $\rho_{XY} = 0$ (Uncorrelated)
  - "When $X$ and $Y$ are uncorrelated, linear estimation is worthless." (What about non-linear?)
  - $\rho_{XY} = 0 \Rightarrow \hat{Y}_{LNH} = m_Y$, $MSE = \sigma_Y^2$

Non-linear estimator: $E(y|X)$

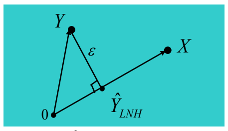
## Linear Homogenous Estimation

This has the form: $\hat{Y}_{LH} = aX$

"$a$" minimizes the MSE: $\frac{d}{da}E[\varepsilon^2] = 0 \Rightarrow a = \frac{E(XY)}{E(X^2)}$

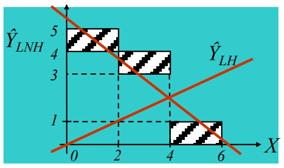
Orthogonality Condition: $E[\varepsilon X] = 0$

$$MSE_{opt} = E[Y^2]\left[1 - \frac{E^2(XY)}{E(X^2)E(Y^2)}\right]$$




Observe that all of this is a special case of $\hat{Y}_{LNH}$ when $m_X = m_Y = 0$

## Earlier Example Cont'd




$MSE_{opt,LNH} = 0.602$

$MSE_{opt,LH} = 10.97$

$\hat{Y}_{LH} = 0.486X$

**Remember**

Linear homogenous estimators are best for zero-mean joint distributions.

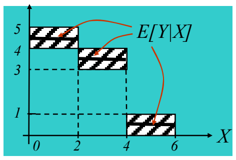
What about nonlinear MSE?

## Nonlinear MSE Estimation

Now we remove the constraint that $\hat{Y}$ must be a linear function of $X$. We will show that the optimal estimator is

$$\hat{Y}_{NL} = E(Y \mid X)$$

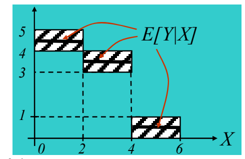
$E(Y \mid X)$ for the previous example is indicated in bold:




$X$ and $Y$ are uniformly distributed over the shaded region.

## Performance

Typically, a double integral is required to calculate the optimal $MSE_{NL}$.




For this example,

$$MSE_{NL} = \int_0^2\int_4^5 (y-4.5)^2 \frac{1}{6}\,dydx + \int_2^4\int_3^4 (y-3.5)^2 \frac{1}{6}\,dydx + \int_4^6\int_0^1 (y-0.5)^2 \frac{1}{6}\,dydx$$

$$= 0.08\overline{3}$$

NL much better

Recall $MSE_{LH}$=$10.97$ and $MSE_{LNH}$=$0.602$.

## Proof That $\hat{Y}_{NL}$=$E[Y|X]$

- The proof includes an interesting use of iterated expectation.
- Begin with $\hat{Y}_{NL} = H(X)$, some arbitrary function of $X$.
- We want $H(X)$ to minimize

just subtract and add it

$$\begin{aligned} MSE_{NL} &= E\{(Y - H(X))^2\} \\ &= E\{[Y - E(Y \mid X) + E(Y \mid X) - H(X)]^2\} \\ &= E\{[Y - E(Y \mid X)]^2\} + 2E\{[Y - E(Y \mid X)][E(Y \mid X) - H(X)]\} \\ &\quad + \{[E(Y \mid X) - H(X)]^2\} \end{aligned}$$

Will address the second term next

## Proof, Cont'd

Use iterated expectation on the second term:

$$E\{[Y - E(Y \mid X)][E(Y \mid X) - H(X)]\}$$

$$= E\{E([Y - E(Y \mid X)]\underbrace{[E(Y \mid X) - H(X)]}) \mid X\}$$

just a function of $X$, so it comes out of the conditional expectation.

$$= E\{\underbrace{E[(Y - E[Y \mid X]) \mid X]}[E(Y \mid X) - H(X)]\}$$

This equals: $E[Y \mid X] - E[Y \mid X] = 0$ so the second term is zero

## Proof, Concluded

The first and third terms remain:

$$MSE_{NL} = \underbrace{E\{[Y - E(Y \mid X)]^2\}} + \underbrace{E\{[E(Y \mid X) - H(X)]^2\}}$$

Ignore this term; it is not affected by $H(X)$.

This is minimized by setting $H(X) = E[Y \mid X]$

$$\therefore \hat{Y}_{NL} = E(Y \mid X)$$

## Nonlinear MSE Estimator for Gaussians

$E(Y \mid X)$ is the mean of $f_{X|Y}(y \mid x)$

Jointly distributed and Gaussian

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

$$= A(x)\exp\underbrace{\left\{-\frac{1}{2(1-\rho_{XY}^2)}\left[B(x) - 2\rho_{XY}\left(\frac{X-\eta_X}{\sigma_X}\right)\left(\frac{Y-\eta_Y}{\sigma_Y}\right) + \left(\frac{Y-\eta_Y}{\sigma_Y}\right)^2\right]\right\}}$$

Exponent is quadratic in $y$; leading term is negative $\Rightarrow f_{Y|X}(y|x)$ is a Gaussian PDF for $y$.

## Because it is Gaussian,

we can find the mean by maximizing $f_{Y|X}(y|x)$, which is equivalent to minimizing the y-dependent portion of the exponent:

$$\left[-2\rho_{XY}\left(\frac{X-\eta_X}{\sigma_X}\right)\left(\frac{Y-\eta_Y}{\sigma_Y}\right)+\left(\frac{Y-\eta_Y}{\sigma_Y}\right)^2\right]$$

The minimization yields

$$\hat{Y}_{NL}=\frac{\sigma_Y\rho_{XY}}{\sigma_X}(X-\eta_X)+\eta_Y$$

LINEAR NON-HOMOGENEOUS!

## Gaussian is Special - Again

**REMEMBER**

The linear non-homogeneous estimator IS THE BEST OF ALL estimators when $X$ and $Y$ are jointly Gaussian.

Random