Lecture 4: Expectation

- Definition and Properties
- Mean and Variance
- Markov and Chebyshev Inequalities
- Covariance and Correlation
- Conditional Expectation
- Iterated Expectation

Expectation

- Let $X \in \mathcal{X}$ be a discrete r.v. with pmf $p_X(x)$ and let $g(x)$ be a function of x. The *expectation* (or *expected value* or *mean*) of $g(X)$ can be defined as

$$\mathrm{E}(g(X)) = \sum_{x \in \mathcal{X}} g(x) p_X(x)$$

- For a continuous r.v. $X \sim f_X(x)$, the expected value of $g(X)$ can be defined as

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

- Properties of expectation:
  - If c is a constant, then $\mathrm{E}(c) = c$
  - Expectation is *linear*, i.e., for any constant a

$$\mathrm{E}[a g_1(X) + g_2(X)] = a\,\mathrm{E}(g_1(X)) + \mathrm{E}(g_2(X))$$

◦ Fundamental Theorem of Expectation: If $Y = g(X) \sim p_Y(y)$, then

$$\mathrm{E}(Y) = \sum_{y \in \mathcal{Y}} y p_Y(y) = \sum_{x \in \mathcal{X}} g(x) p_X(x) = \mathrm{E}(g(X))$$

The corresponding formula for $f_Y(y)$ uses integrals instead of sums:

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y)\, dy$$

Proof: We prove the theorem for discrete r.v.s. Consider

$$\begin{aligned}
\mathrm{E}(Y) &= \sum_y y p_Y(y) \\
&= \sum_y y \sum_{\{x:\, g(x)=y\}} p_X(x) \\
&= \sum_y \sum_{\{x:\, g(x)=y\}} y p_X(x) = \sum_y \sum_{\{x:\, g(x)=y\}} g(x) p_X(x) = \sum_x g(x) p_X(x)
\end{aligned}$$

Thus $\mathrm{E}(Y) = \mathrm{E}(g(X))$ can be found using either $f_X(x)$ or $f_Y(y)$

It is often much easier to use $f_X(x)$ than to first find $f_Y(y)$ then find $\mathrm{E}(Y)$

- Remark: We know that a r.v. is completely specified by its cdf (pdf, pmf), so why do we need expectation?
  - Expectation provides a *summary* or an *estimate* of the r.v. — a single number — instead of specifying the entire distribution.
  - It is far easier to estimate the expectation of a r.v. from data than to estimate its distribution
  - Expectation can be used to bound or estimate probabilities of interesting events (as we shall see)

Mean and Variance

- The *first moment* (or *mean*) of $X \sim f_X(x)$ is the expectation

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

- The *second moment* (or *mean square* or *average power*) of X is

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx$$

- The *variance* of X is

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \\
&= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}(X) + (\mathrm{E}(X))^2\right] \\
&= \mathrm{E}(X^2) - 2(\mathrm{E}(X))^2 + (\mathrm{E}(X))^2 \\
&= \mathrm{E}(X^2) - (\mathrm{E}(X))^2
\end{aligned}$$

- The *standard deviation* of X is defined as $\sigma_X = \sqrt{\mathrm{Var}(X)}$, i.e., $\mathrm{Var}(X) = \sigma^2$

Mean and Variance for Famous RVs

Random Variable	Mean	Variance
$\mathrm{Bern}(p)$	p	$p(1-p)$
$\mathrm{Geom}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\mathrm{Binom}(n,p)$	np	$np(1-p)$
$\mathrm{Poisson}(\lambda)$	λ	λ
$\mathrm{U}[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathrm{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2

Expectation Might not Exist

- Expectation can be infinite. For example

$$f_X(x) = \begin{cases} 1/x^2 & 1 \le x < \infty \\ 0 & \text{otherwise} \end{cases} \Rightarrow \; \mathrm{E}(X) = \int_1^\infty x/x^2 \, dx = \infty$$

- Expectation may not exist. To find conditions for expectation to exist, consider

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = -\int_{-\infty}^{0} |x| f_X(x) \, dx + \int_0^\infty |x| f_X(x) \, dx \,,$$

so either $\int_{-\infty}^{0} |x| f_X(x) \, dx$ or $\int_0^\infty |x| f_X(x) \, dx$ must be finite

- Example: the *standard Cauchy* r.v. has the pdf

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Since both $\int_{-\infty}^{0} |x| f(x) \, dx$ and $\int_0^\infty |x| f(x) \, dx$ are infinite, its mean does not exist! (The second moment of the Cauchy is $\mathrm{E}(X^2) = \infty$, so it exists)

Bounding Probability Using Expectation

- In many cases we do not know the distribution of a r.v. X but want to find the probability of an event such as $\{X > a\}$ or $\{|X - \mathrm{E}(X)| > a\}$
- The Markov and Chebyshev inequalities give bounds on the probabilities of such events in terms of the mean and variance of the random variable
- Example: Let $X \geq 0$ represent the age of a person in the Bay Area. If we know that $\mathrm{E}(X) = 35$ years, what fraction of the population is ≥ 70 years old?

  Clearly we cannot answer this question knowing only the mean, but we can say that $\mathrm{P}\{X \geq 70\} \leq 0.5$, since otherwise the mean would be larger than 35
- This is an application of the *Markov inequality*

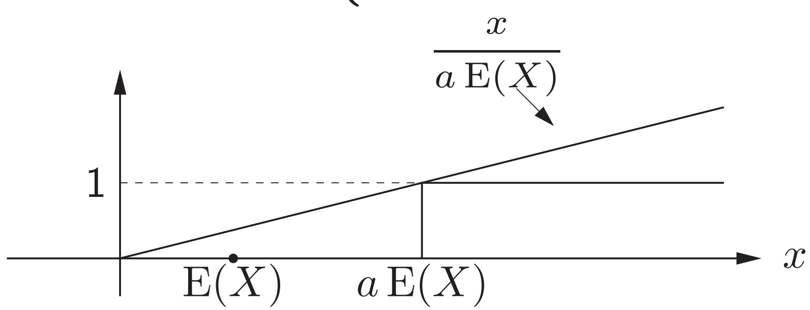
Markov Inequality

- For any r.v. $X \geq 0$ with finite mean $\mathrm{E}(X)$ and any $a > 1$,

$$\mathrm{P}\{X \geq a\,\mathrm{E}(X)\} \leq \frac{1}{a}$$

Proof: Define the *indicator function* of the set $A = \{x \geq a\,\mathrm{E}(X)\}$:

$$\mathrm{I}_A(x) = \begin{cases} 1 & x \geq a\,\mathrm{E}(X) \\ 0 & \text{otherwise} \end{cases}$$



Clearly, $\mathrm{I}_A \leq \dfrac{X}{a\,\mathrm{E}(X)}$

Since $\mathrm{E}(\mathrm{I}_A) = \mathrm{P}(A) = \mathrm{P}\{X \geq a\,\mathrm{E}(X)\}$, taking the expectations of both sides we obtain the Markov Inequality

- The Markov inequality can be *very* loose. If $X \sim \mathrm{Exp}(1)$, then

$$\mathrm{P}\{X \geq 10\} = e^{-10} \approx 4.54 \times 10^{-5}$$

  The Markov inequality gives

$$\mathrm{P}\{X \geq 10\} \leq \tfrac{1}{10},$$

  which is very pessimistic

- But, it is the *tightest* possible inequality on $\mathrm{P}\{X \geq a\,\mathrm{E}(X)\}$ when we are given only the mean of X

  To show this, note that the inequality is tight for the following r.v.:

$$X = \begin{cases} a\,\mathrm{E}(X) & \text{with probability } 1/a \\ 0 & \text{with probability } 1 - 1/a \end{cases}$$

Chebyshev Inequality

- Let X be a device parameter in an integrated circuit (IC) with known mean and variance. The IC is out-of-spec if X is more than, say, $3\sigma_X$ away from its mean. We wish to find the fraction of out-of-spec ICs, namely, $\mathrm{P}\{|X - \mathrm{E}(X)| \geq 3\sigma_X\}$

  The *Chebyshev inequality* gives us an upper bound on this fraction in terms the mean and variance of X

- Let X be a r.v. with known $\mathrm{E}(X)$ and $\mathrm{Var}(X) = \sigma_X^2$. The Chebyshev inequality states that for every $a > 1$,

$$\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\sigma_X\} \leq \frac{1}{a^2}$$

  Proof: We use the Markov inequality. Define the r.v. $Y = (X - \mathrm{E}(X))^2 \geq 0$. Since $\mathrm{E}(Y) = \sigma_X^2$, the Markov inequality gives

$$\mathrm{P}\{Y \geq a^2\sigma_X^2\} \leq \frac{1}{a^2}$$

  But $\{|X - \mathrm{E}(X)| \geq a\sigma_X\}$ occurs iff $\{Y \geq a^2\sigma_X^2\}$. Thus

$$\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\sigma_X\} \leq \frac{1}{a^2}$$

- The Chebyshev inequality can be very loose. Let $X \sim \mathcal{N}(0,1)$. Using the Chebyshev inequality we obtain

$$\mathrm{P}\{|X| \geq 3\} \leq \tfrac{1}{9},$$

which is very pessimistic compared to the actual value $2Q(3) \approx 2 \times 10^{-3}$

- But, it is the tightest inequality on $\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\sigma_X\}$ given knowledge only of the mean and variance of X. To show this, note that equality is achieved for the random variable

$$X = \begin{cases} \mathrm{E}(X) + a\sigma_X & \text{with probability } 1/2a^2 \\ \mathrm{E}(X) - a\sigma_X & \text{with probability } 1/2a^2 \\ \mathrm{E}(X) & \text{with probability } 1 - 1/a^2 \end{cases}$$

Expectation Involving Two RVs

- Let $(X, Y) \sim f_{X,Y}(x, y)$ and let $g(x, y)$ be a function of x and y. The expectation of $g(X, Y)$ is given by

$$\mathrm{E}(g(X, Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy$$

  The function $g(X, Y)$ may be X, Y, X^2, $X + Y$, etc.

- The *correlation* of X and Y is defined as $\mathrm{E}(XY)$
- The *covariance* of X and Y is defined as

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathrm{E}\left[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right] \\ &= \mathrm{E}\left[XY - X\,\mathrm{E}(Y) - Y\,\mathrm{E}(X) + \mathrm{E}(X)\,\mathrm{E}(Y)\right] \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) \end{aligned}$$

- Note that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

- Example: Find $\mathrm{E}(X)$, $\mathrm{Var}(X)$, and $\mathrm{Cov}(X,Y)$ for $(X,Y) \sim f(x,y)$ where

$$f(x,y) \;=\; \begin{cases} 2 & x \geq 0,\, y \geq 0,\, x+y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution:

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,dx\,dy \\ &= \int_0^1\int_0^{1-x} 2x\,dy\,dx \;=\; 2\int_0^1 (1-x)x\,dx \;=\; 2(\tfrac{1}{2}-\tfrac{1}{3}) \;=\; \tfrac{1}{3} \end{aligned}$$

Since $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$, we need to find the second moment:

$$\mathrm{E}(X^2) \;=\; 2\int_0^1 (1-x)x^2\,dx \;=\; 2(\tfrac{1}{3}-\tfrac{1}{4}) \;=\; \tfrac{1}{6}\,,$$

hence

$$\mathrm{Var}(X) \;=\; \tfrac{1}{6} - (\tfrac{1}{9})^2 \;=\; \tfrac{1}{6} - \tfrac{1}{9} \;=\; \tfrac{1}{18}$$

By symmetry, $\mathrm{E}(Y) = \mathrm{E}(X) = \frac{1}{3}$. Thus the covariance of X and Y is

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= 2\int_0^1 \int_0^{1-x} xy\,dy\,dx - \mathrm{E}(X)\,\mathrm{E}(Y) \\
&= \int_0^1 x(1-x)^2\,dx - \tfrac{1}{9} \;=\; \tfrac{1}{12} - \tfrac{1}{9} = -\tfrac{1}{36}
\end{aligned}$$

Uncorrelation vs. Independence

- X and Y are said to be *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$
- If X and Y are independent then they are uncorrelated, since

$$\begin{aligned}
\mathrm{E}(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x, y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} y f(y) \left(\int_{-\infty}^{\infty} x f_X(x)\, dx \right) dy \\
&= \mathrm{E}(X) \int_{-\infty}^{\infty} y f(y)\, dy \;=\; \mathrm{E}(X)\, \mathrm{E}(Y)
\end{aligned}$$

  Therefore $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\, \mathrm{E}(Y) = 0$

- X and Y uncorrelated does *not* necessarily imply that they are independent

- Example: Let $X, Y \in \{-2, -1, +1, +2\}$ such that

$$p(x,y) = \begin{cases} \frac{2}{5} & (x,y) = (+1,+1), (-1,-1) \\ \frac{1}{10} & (x,y) = (+2,-2), (-2,+2) \\ 0 & \text{otherwise} \end{cases}$$

Are X and Y independent? Are they uncorrelated?

- Solution: Clearly X and Y are not independent

Let's check their covariance:

$$\mathrm{E}(X) = \frac{2}{5} - \frac{2}{5} - \frac{2}{10} + \frac{2}{10} = 0$$

$$\mathrm{E}(Y) = 0 \quad \text{(by symmetry)}$$

$$\mathrm{E}(XY) = \frac{2}{5} + \frac{2}{5} - \frac{4}{10} - \frac{4}{10} = 0$$

Thus, $\mathrm{Cov}(X,Y) = 0$, and X and Y are uncorrelated!

Correlation Coefficient

- The *correlation coefficient* of X and Y is defined as

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- Fact: $|\rho_{X,Y}| \leq 1$. To show this consider

$$\mathrm{E}\left[\left(\frac{X-\mathrm{E}(X)}{\sigma_X} \pm \frac{Y-\mathrm{E}(Y)}{\sigma_Y}\right)^2\right] \geq 0$$

$$\frac{\mathrm{E}\left[(X-\mathrm{E}(X))^2\right]}{\sigma_X^2} + \frac{\mathrm{E}\left[(Y-\mathrm{E}(Y))^2\right]}{\sigma_Y^2} \pm 2\frac{\mathrm{E}\left[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right]}{\sigma_X\sigma_Y} \geq 0$$

$$1+1\pm 2\rho_{X,Y} \geq 0 \;\Rightarrow\; -2 \leq 2\rho_{X,Y} \leq 2 \;\Rightarrow\; |\rho_{X,Y}| \leq 1$$

- From the proof, we see that $\rho_{X,Y} = \pm 1$ iff $\dfrac{X-\mathrm{E}(X)}{\sigma_X} = \pm\dfrac{Y-\mathrm{E}(Y)}{\sigma_Y}$ (equality with probability 1), i.e., iff $X-\mathrm{E}(X)$ is a linear function of $Y-\mathrm{E}(Y)$.

- Note: We shall see that $\rho_{X,Y}$ is a measure of how closely $X-\mathrm{E}(X)$ can be approximated or estimated by a linear function of $Y-\mathrm{E}(Y)$

Conditional Expectation

- Conditioning on an event: Let $X \sim p_X(x)$ be a r.v. and A be a nonzero probability event. We can define the conditional pmf of X given $X \in A$ as

$$p_{X|A}(x) = \mathrm{P}\{X = x \,|\, X \in A\} = \frac{\mathrm{P}\{X = x, X \in A\}}{\mathrm{P}\{X \in A\}} = \begin{cases} \frac{p_X(x)}{\mathrm{P}\{X \in A\}} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

  Note that $p_{X|A}(x)$ is a pmf on X

- Similarly, if $X \sim f_X(x)$,

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathrm{P}\{X \in A\}} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

  is a pdf on X

- Example: Let $X \sim \mathrm{Exp}(\lambda)$ and $A = \{X > a\}$, for some $a > 0$. The conditional pdf of X given A is $\lambda e^{-\lambda(x-a)}$, for $a \geq 0$, and 0 otherwise

- We define the *conditional expectation* of $g(X)$ given $X \in A$ as

$$\mathrm{E}(g(X) \mid A) = \int_{-\infty}^{\infty} g(x) f_{X|A}(x)\, dx$$

- Example: Find $\mathrm{E}(X \mid A)$ and $\mathrm{E}(X^2 \mid A)$ for the previous example

- Total Expectation Theorem: Let $X \sim f_X(x)$ and $A_1, A_2, \ldots, A_n$ be disjoint nonzero probability events with $\mathrm{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathrm{P}(A_i) = 1$. Then

$$\mathrm{E}(g(X)) = \sum_{i=1}^{n} \mathrm{P}\{X \in A_i\}\, \mathrm{E}(g(X)|A_i)$$

  Proof: First note that by the law of total probability

$$f_X(x) = \sum_{i=1}^{n} \mathrm{P}(A_i) f_{X|A_i}(x)$$

$$
\begin{aligned}
\mathrm{E}(g(X)) &= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} g(x) \sum_{i=1}^{n} \mathrm{P}(A_i) f_{X|A_i}(x)\, dx \\
&= \sum_{i=1}^{n} \mathrm{P}(A_i) \int_{-\infty}^{\infty} g(x) f_{X|A_i}(x)\, dx = \sum_{i=1}^{n} \mathrm{P}(A_i)\, \mathrm{E}(g(X) \mid A_i)
\end{aligned}
$$

This result is useful in computing expectation by *divide-and-conquer*

- Example: Mean and variance of piecewise uniform pdf. Let X be a continuous r.v. with the piecewise uniform pdf

$$
f_X(x) = \begin{cases} 1/3 & \text{if } 0 \le x \le 1 \\ 2/3 & \text{if } 1 < x \le 2 \\ 0 & \text{otherwise} \end{cases}
$$

Find the mean and variance of X

Solution: The events $A_1 = \{X \in [0,1]\}$ and $A_2 = \{X \in (1,2]\}$ are disjoint and the sum of their probabilities is 1. The mean and second moment of X can be expressed as

$$\mathrm{E}(X) = \sum_{i=1}^{2} \mathrm{P}\{X \in A_i\}\, \mathrm{E}(X \mid A_i) = \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{3}{2} = \frac{7}{6}$$

$$\mathrm{E}(X^2) = \sum_{i=1}^{2} \mathrm{P}\{X \in A_i\}\, \mathrm{E}(X^2 \mid A_i) = \frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{7}{3} = \frac{15}{9}$$

Thus

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{11}{36}$$

Conditioning on a RV

- Let $(X, Y) \sim f_{X,Y}(x, y)$. If $f_Y(y) \neq 0$, the *conditional pdf* of X given $Y = y$ is given by

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

- We know that $f_{X|Y}(x|y)$ is a pdf for X (function of y), so we can define the expectation of any function $g(X, Y)$ w.r.t. $f_{X|Y}(x|y)$ as

$$\mathrm{E}(g(X, Y) \mid Y = y) = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y)\, dx$$

- Example: If $g(X, Y) = X$, then the conditional expectation of X given $Y = y$ is

$$\mathrm{E}(X \mid Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$

- Example: If $g(X, Y) = XY$, then $\mathrm{E}(XY \mid Y = y) = y\, \mathrm{E}(X \mid Y = y)$

- Example: Let

$$f_{X,Y}(x,y) = \begin{cases} 2 & \text{if } x \ge 0,\ y \ge 0,\ x + y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $\mathrm{E}(X \mid Y = y)$ and $\mathrm{E}(XY \mid Y = y)$

Solution: We already know that

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{1-y} & \text{if } x \ge 0,\ y \ge 0,\ x + y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus

$$\begin{aligned} \mathrm{E}(X|Y = y) &= \int_0^{1-y} \frac{1}{1-y} x \, dx \\ &= \frac{1-y}{2}, \quad 0 \le y < 1 \end{aligned}$$

Now to find $\mathrm{E}(XY \mid Y = y)$, note that

$$\begin{aligned} \mathrm{E}(XY \mid Y = y) &= y \, \mathrm{E}(X \mid Y = y) \\ &= \frac{y(1-y)}{2}, \quad 0 \le y < 1 \end{aligned}$$

Conditional Expectation as a RV

- We define the *conditional expectation* of $g(X,Y)$ given Y as the random variable $\mathrm{E}(g(X,Y)\,|\,Y)$, which is a function of the random variable Y
- In particular, $\mathrm{E}(X\,|\,Y)$ is the conditional expectation of X given Y, a r.v. that is a function of Y
- Example (continuation of previous example): Find the pdf of $\mathrm{E}(X\,|\,Y)$

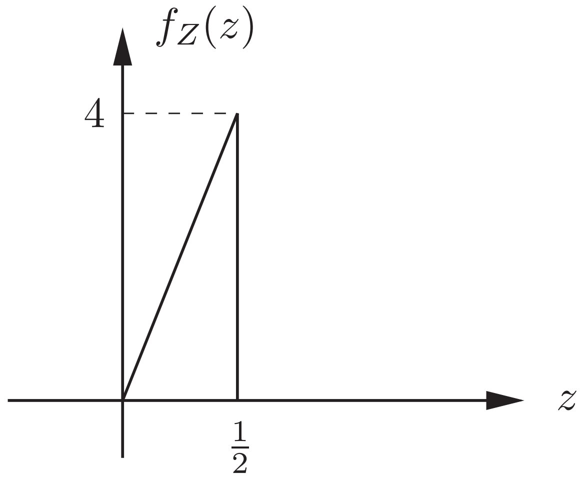
  Solution: The conditional expectation of X given Y is the r.v.

  $$\mathrm{E}(X\,|\,Y) = \frac{1-Y}{2} \triangleq Z$$

  The pdf of Z is given by

  $$f_Z(z) = 8z\,, \quad 0 < z \le \tfrac{1}{2}$$

Graph of $f_Z(z)$:



Now let's find the expected value of the r.v. Z

$$\mathrm{E}(Z) = \int_0^{\frac{1}{2}} 8z^2 dz = \tfrac{1}{3} = E(X)$$

i.e., for this example $\mathrm{E}[E(X \mid Y)] = \mathrm{E}(X)$. This is in fact true for any X and Y

Iterated Expectation

- In general we can find $\mathrm{E}(g(X,Y))$ using *iterated expectation* as

$$\mathrm{E}(g(X,Y)) \;=\; \mathrm{E}_Y\left[E_X(g(X,Y)\,|\,Y)\right],$$

where E_X means expectation w.r.t. $f_{X|Y}(x|y)$ and E_Y means expectation w.r.t. $f_Y(y)$. To show this, consider

$$\begin{aligned}
\mathrm{E}_Y\left[\mathrm{E}_X(g(X,Y)\,|\,Y)\right] &= \int_{-\infty}^{\infty} \mathrm{E}_X(g(X,Y)\,|\,Y=y) f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y)\,dx\right) f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y) f_Y(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy = \mathrm{E}(g(X,Y))
\end{aligned}$$

- This result can be very useful in computing expectation

- Example: A coin has random bias $P \in [0,1]$ with pdf $f_P(p) = 2(1-p)$. The coin is flipped n times. Let N be the number of heads. Find $\mathrm{E}(N)$

  Solution: Of course, we could first find the pmf of N, then find its expectation. Using iterated expectation we can find N more easily

$$\begin{aligned}
\mathrm{E}(N) &= \mathrm{E}_P[\mathrm{E}_N(N \mid P)] \\
&= \mathrm{E}_P(nP) \\
&= n \int_0^1 2(1-p)p\,dp \;=\; \tfrac{1}{3}n
\end{aligned}$$

- Example: Let $\mathrm{E}(X \mid Y) = Y^2$ and $Y \sim \mathrm{U}[0,1]$. Find $\mathrm{E}(X)$

  Solution: We cannot first find the pdf of X, since we do not know $f_{X|Y}(x|y)$, but using iterated expectation we can easily find

$$\mathrm{E}(X) = \mathrm{E}_Y(\mathrm{E}_X(X \mid Y)) \;=\; \int_0^1 y^2\,dy \;=\; \tfrac{1}{3}$$

Conditional Variance

- Let X and Y be two r.v.s. We define the *conditional variance* of X given $Y = y$ to be the variance of X using $f_{X|Y}(x|y)$, i.e.,

$$\begin{aligned}\mathrm{Var}(X \mid Y = y) &= \mathrm{E}\left[(X - \mathrm{E}(X \mid Y = y))^2 \mid Y = y\right] \\ &= \mathrm{E}(X^2 \mid Y = y) - [\mathrm{E}(X \mid Y = y)]^2\end{aligned}$$

- The r.v. $\mathrm{Var}(X \mid Y)$ is simply a function of Y that takes on the values $\mathrm{Var}(X \mid Y = y)$. Its expected value is

$$\mathrm{E}_Y\left[\mathrm{Var}(X \mid Y)\right] = \mathrm{E}_Y\left[\mathrm{E}(X^2 \mid Y) - (\mathrm{E}(X \mid Y))^2\right] = \mathrm{E}(X^2) - \mathrm{E}\left[(\mathrm{E}(X \mid Y))^2\right]$$

- Since $\mathrm{E}(X \mid Y)$ is a r.v., it has a variance

$$\mathrm{Var}(\mathrm{E}(X \mid Y)) = \mathrm{E}_Y\left[\left(\mathrm{E}(X \mid Y) - \mathrm{E}[\mathrm{E}(X \mid Y)]\right)^2\right] = \mathrm{E}\left[(\mathrm{E}(X \mid Y))^2\right] - (\mathrm{E}(X))^2$$

- *Law of Conditional Variances*: Adding the above expressions, we obtain

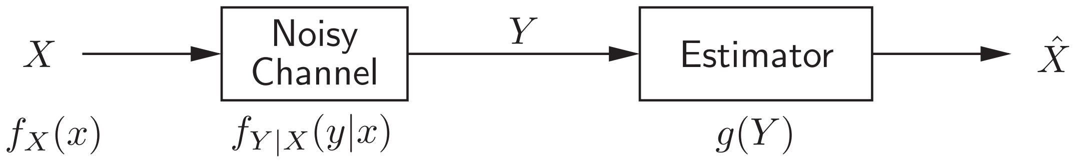
$$\mathrm{Var}(X) = \mathrm{E}\left(\mathrm{Var}(X \mid Y)\right) + \mathrm{Var}\left(\mathrm{E}(X \mid Y)\right)$$

Lecture 5: Mean Square Error Estimation

- Minimum MSE Estimation
- Linear Estimation
- Jointly Gaussian Random Variables

Minimum MSE Estimation

- Consider the following signal processing problem:



- X is a signal with known statistics, i.e., known pdf $f_X(x)$
- The signal is transmitted (or stored) over a noisy channel with known statistics, i.e., conditional pdf $f_{Y|X}(y|x)$
- We observe the signal Y and wish to find the *estimate* $\hat{X} = g(Y)$ of X that minimizes the *mean square error*

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

- The $\hat{X}$ that achieves the minimum MSE is called the *minimum MSE estimate* (MMSE) of X (given Y)

MMSE Estimate

- Theorem: The MMSE estimate of X given the observation Y and complete knowledge of the joint pdf $f_{X,Y}(x,y)$ is

$$\hat{X} = \mathrm{E}(X \mid Y)\,,$$

  and the MSE of $\hat{X}$, i.e., the minimum MSE, is

$$\mathrm{MMSE} = \mathrm{E}_Y(\mathrm{Var}(X \mid Y)) = \mathrm{E}(X^2) - \mathrm{E}\left[(\mathrm{E}(X \mid Y))^2\right]$$

- Properties of the minimum MSE estimator:
  - Since $\mathrm{E}(\hat{X}) = \mathrm{E}_Y[\mathrm{E}(X \mid Y)] = \mathrm{E}(X)$, the best MSE estimate is *unbiased*
  - If X and Y are independent, then the best MSE estimate is $\mathrm{E}(X)$
  - The conditional expectation of the estimation error, $\mathrm{E}\left[(X - \hat{X}) \mid Y = y\right]$, is 0 for all y, i.e., the error is unbiased for every $Y = y$

- The estimation error and the estimate are "orthogonal"

$$
\begin{aligned}
\mathrm{E}\left[(X-\hat{X})\hat{X}\right] &= \mathrm{E}_Y\left[\mathrm{E}\left((X-\hat{X})\hat{X} \mid Y\right)\right] \\
&= \mathrm{E}_Y\left[\hat{X}\,\mathrm{E}((X-\hat{X}) \mid Y)\right] \\
&= \mathrm{E}_Y\left[\hat{X}(\mathrm{E}(X \mid Y)-\hat{X}) \mid Y)\right] \\
&= 0
\end{aligned}
$$

  In fact, the estimation error is orthogonal to *any* function $g(Y)$ of Y

- From the law of conditional variance

$$
\mathrm{Var}(X) = \mathrm{Var}(\hat{X}) + \mathrm{E}(\mathrm{Var}(X \mid Y)),
$$

  i.e., the sum of the variance of the estimate and the minimum MSE is equal to the variance of the signal

- Proof of Theorem: We first show that $\min_a \mathrm{E}\left((X-a)^2\right) = \mathrm{Var}(X)$ and that the minimum is achieved for $a = \mathrm{E}(X)$, i.e., in the absence of any observations, the mean of X is its minimum MSE estimate

  To show this, consider

$$\begin{aligned}
\mathrm{E}\left[(X-a)^2\right] &= \mathrm{E}\left[(X-\mathrm{E}(X) \,+\, \mathrm{E}(X)-a)^2\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \,+\, \big(\mathrm{E}(X)-a\big)^2 + \\
&\qquad 2\,\mathrm{E}(X-\mathrm{E}(X))(\mathrm{E}(X)-a) \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \,+\, \big(\mathrm{E}(X)-a\big)^2 \\
&\geq \mathrm{E}\left[(X-\mathrm{E}(X))^2\right]
\end{aligned}$$

  Equality holds if and only if $a = \mathrm{E}(X)$

  We use this result to show that $\mathrm{E}(X \mid Y)$ is the MMSE estimate of X given Y.

First write

$$\mathrm{E}\left[(X - g(Y))^2\right] \;=\; \mathrm{E}_Y\left[\mathrm{E}_X((X - g(Y))^2 \,|\, Y)\right]$$

From the previous result we know that for each $Y = y$ the minimum value for $\mathrm{E}_X\left[(X - g(y))^2 \,|\, Y = y\right]$ is obtained when $g(y) = \mathrm{E}(X \,|\, Y = y)$

Therefore the overall MSE is minimized for $g(Y) = \mathrm{E}(X \,|\, Y)$

In fact, $\mathrm{E}(X \,|\, Y)$ minimizes the MSE conditioned on every $Y = y$ and not just its average over Y

To find the minimum MSE, consider

$$\mathrm{E}\left[(X - \mathrm{E}(X \,|\, Y))^2\right] = \mathrm{E}_Y \, \mathrm{E}_X\left[(X - \mathrm{E}(X \,|\, Y))^2 \,|\, Y\right] = \mathrm{E}_Y \, \mathrm{Var}(X \,|\, Y)$$

Example

- Again let

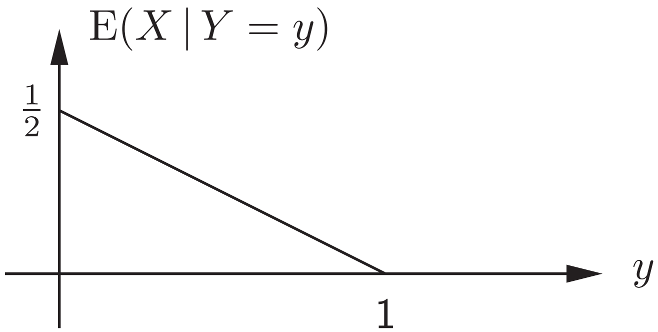
$$f_{X,Y}(x,y) = \begin{cases} 2 & x \geq 0,\ y \geq 0,\ x+y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the MMSE estimate of X given Y and its MSE

Solution: We know that

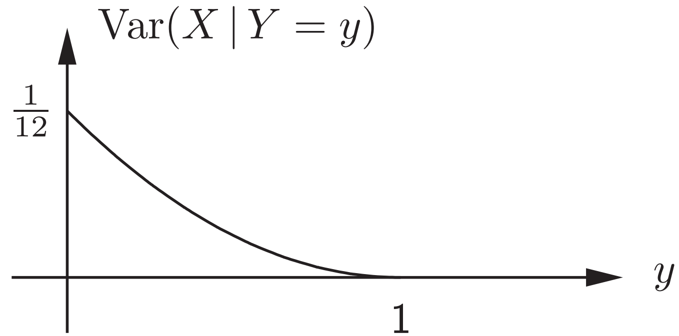
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{1}{1-y} & 0 \leq y \leq 1,\ 0 \leq x \leq 1-y \\ 0 & \text{otherwise} \end{cases}$$

Thus the MMSE estimate is given by $\mathrm{E}(X \mid Y) = \dfrac{1-Y}{2}, \quad 0 \leq Y \leq 1$



And, for $Y = y$, the minimum MSE is given by

$$\mathrm{Var}(X \mid Y = y) = \frac{(1-y)^2}{12}, \; 0 \le y < 1$$



Thus the minimum MSE is $\mathrm{E}_Y(\mathrm{Var}(X \mid Y)) = \frac{1}{24}$, compared to $\mathrm{Var}(X) = \frac{1}{18}$.

The difference is $\mathrm{Var}(\mathrm{E}(X \mid Y)) = \frac{1}{72}$, i.e., the variance of the estimate

Additive Gaussian Noise Channel

- Consider a communication channel with input $X \sim \mathcal{N}(\mu, P)$, noise $Z \sim \mathcal{N}(0, N)$, and output $Y = X + Z$. X and Z are independent

  Find the MMSE estimate of X given Y and its MSE, i.e., $\mathrm{E}(X \mid Y)$ and $\mathrm{E}(\mathrm{Var}(X \mid Y))$

- To find $f_{X|Y}(x|y)$ we use Bayes rule:

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

  To find $f_{Y|X}(y|x)$, we use "my" favorite trick: since Y is the sum of two independent r.v.s,

$$f_{Y|X}(y|x) = f_{Z|X}(y - x|x) = f_Z(y - x) = \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-x)^2}{2N}}$$

  In other words, $Y \mid \{X = x\} \sim \mathcal{N}(x, N)$

- Since X and Z are independent and Gaussian, $Y = X + Z \sim \mathcal{N}(\mu,\, P+N)$. Thus

$$f_{X|Y}(x|y) \;=\; \frac{1}{\sqrt{2\pi \frac{PN}{P+N}}} e^{-\frac{\left(x - (\frac{P}{P+N}y + \frac{N}{P+N}\mu)\right)^2}{2\frac{PN}{P+N}}}$$

$$X \,|\, \{Y=y\} \sim \mathcal{N}\left(\frac{P}{P+N}y + \frac{N}{P+N}\mu\,,\ \frac{PN}{P+N}\right)$$

$$\mathrm{E}(X \,|\, Y) = \frac{P}{P+N}Y + \frac{N}{P+N}\mu\,, \quad \mathrm{E}(\mathrm{Var}(X \,|\, Y)) = \frac{PN}{P+N}$$

- Note: In the above two examples, the MMSE estimate turned out to be an affine function of Y (i.e., of the form $aY + b$)

  This is not always the case; for example, let

$$f(x|y) \;=\; \begin{cases} ye^{-yx} & x \ge 0,\ y > 0 \\ 0 & \text{otherwise} \end{cases}$$

  In this case $\mathrm{E}(X \,|\, Y) = 1/Y$

Linear Estimation

- To find the MMSE estimate one needs to know the statistics of the signal and the channel — $f_{X,Y}(x,y)$ — which is rarely the case in practice
- We typically have estimates only of the first and second moments of the signal and the observation, i.e., means, variances, and covariance of X and Y
- This is not, in general, sufficient information for computing the MMSE estimate, but as we shall see is enough to compute the MMSE linear (or affine) estimate of the signal X given the observation Y, i.e., the estimate of the form

  $$\hat{X} = aY + b$$

  that minimizes the mean square error

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

MMSE Linear Estimate

- Theorem: The MMSE linear estimate of X given Y is

$$\begin{aligned}\hat{X} &= \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2}\big(Y - \mathrm{E}(Y)\big) \;+\; \mathrm{E}(X) \\ &= \rho_{X,Y}\sigma_X\left(\frac{Y - \mathrm{E}(Y)}{\sigma_Y}\right) \;+\; \mathrm{E}(X)\end{aligned}$$

  and its MSE is given by

$$\mathrm{MSE} \;=\; \sigma_X^2 - \frac{\mathrm{Cov}^2(X,Y)}{\sigma_Y^2} \;=\; (1-\rho_{X,Y}^2)\sigma_X^2$$

- Properties of best linear MSE estimate:
  - $\mathrm{E}(\hat{X}) = \mathrm{E}(X)$, i.e., estimate is unbiased (also true for best MSE estimate)
  - If $\rho_{X,Y} = 0$, i.e., X and Y are uncorrelated, then $\hat{X} = \mathrm{E}(X)$ — the observation Y is ignored!
  - If $\rho_{X,Y} = \pm 1$, i.e., $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$ are linearly dependent, then the linear estimate is perfect

- Proof: To find the coefficients a and b we take derivatives and set them to 0

$$\mathrm{MSE} \;=\; \mathrm{E}\left[(X-\hat{X})^2\right] \;=\; \mathrm{E}\left[(X-(aY+b))^2\right]$$

$$\frac{\partial}{\partial b}\mathrm{MSE} \;=\; 0 \;\Rightarrow\; \mathrm{E}(X-\hat{X})=0 \;\Rightarrow\; \mathrm{E}(\hat{X})=\mathrm{E}(X)$$

$$\frac{\partial}{\partial a}\mathrm{MSE} \;=\; 0 \;\Rightarrow\; \mathrm{E}\left[(X-\hat{X})Y\right]=0$$

Thus

$$\mathrm{E}\left[[(X-\mathrm{E}(X))-(\hat{X}-\mathrm{E}(\hat{X}))]\cdot[Y-\mathrm{E}(Y)]\right] \;=\; 0$$

or

$$\mathrm{E}\left[[(X-\mathrm{E}(X))-a(Y-\mathrm{E}(Y))]\cdot[Y-\mathrm{E}(Y)]\right] \;=\; 0$$

hence

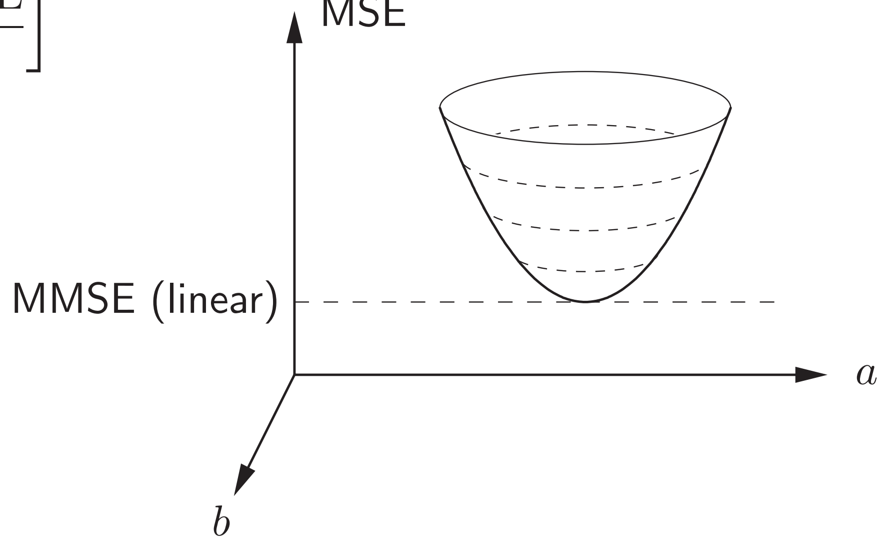
$$\mathrm{Cov}(X,Y)-a\sigma_Y^2 \;=\; 0$$

Therefore

$$a \;=\; \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2} \quad \text{and} \quad b \;=\; \mathrm{E}(X)-\frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2}\mathrm{E}(Y)$$

Note: These a and b *globally* minimize the MSE since the MSE is *convex* in a and b, which can be established by showing that the *Hessian* matrix is *nonnegative definite* (check it)

$$\mathcal{H} = \begin{bmatrix} \dfrac{\partial^2 \mathrm{MSE}}{\partial a^2} & \dfrac{\partial^2 \mathrm{MSE}}{\partial a \partial b} \\[2ex] \dfrac{\partial^2 \mathrm{MSE}}{\partial a \partial b} & \dfrac{\partial^2 \mathrm{MSE}}{\partial b^2} \end{bmatrix}$$



To find the MMSE, substitute a and b into the MSE expression $\mathrm{E}\left[(X - (aY + b))^2\right]$

Geometric Formulation of Linear Estimation

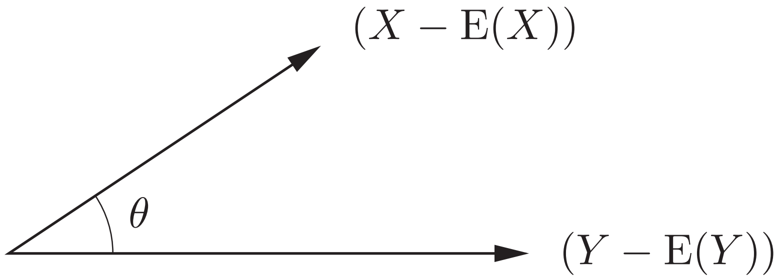
- First we introduce some needed background
- A *vector space* $\mathcal{V}$, e.g., Euclidean space, consists of a set of vectors that are closed under two operations:
  - *vector addition*: if $v_1, v_2 \in \mathcal{V}$ then $v_1 + v_2 \in \mathcal{V}$
  - *scalar multiplication*: if $a \in \mathbf{R}$ and $v \in \mathcal{V}$, then $av \in \mathcal{V}$
- An *inner product*, e.g., dot product in Euclidean space, is a real-valued operation $u \cdot v$ satisfying these three conditions:
  - commutativity: $u \cdot v = v \cdot u$
  - linearity: $(au + v) \cdot w = a(u \cdot w) + v \cdot w$
  - nonnegativity: $u \cdot u \geq 0$ and $u \cdot u = 0$ iff $u = 0$
- The *norm* of u is defined as $\|u\| = \sqrt{u \cdot u}$
- u and v are *orthogonal* (written $u \perp v$) if $u \cdot v = 0$
- A vector space with an inner product is called an *inner product space*. Example: Euclidean space with dot product

- Now let's go back to linear estimation
- View $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$ as vectors in an inner product space

  This inner product space $\mathcal{V}$ consists of all zero mean random variables defined over the same probability space, with

  - vector addition: $V_1 + V_2 \in \mathcal{V}$
    adding two zero mean r.v.s yields a zero mean r.v.
  - scalar multiplication: $aV \in \mathcal{V}$
    multiplying a zero mean r.v. by a constant yields a zero mean r.v.
  - inner product: $\mathrm{E}(V_1 V_2)$
    exercise: check that this is a legitimate inner product
  - norm of V: $\|V\| = \sqrt{\mathrm{E}(V^2)} = \sigma_V$

- So we have the following picture for the r.v.s $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$:

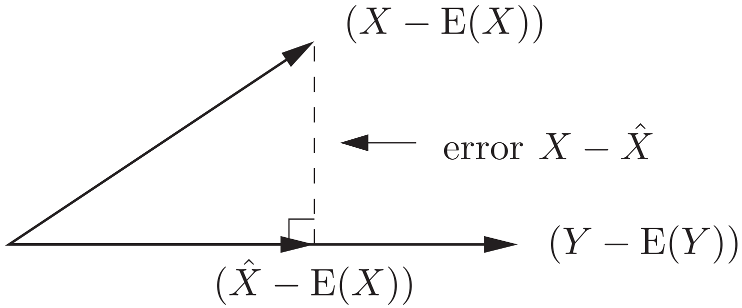


inner product	$\Leftrightarrow$	$\mathrm{Cov}(X, Y)$
norm of $(X - \mathrm{E}(X))$	$\Leftrightarrow$	σ_X
norm of $(Y - \mathrm{E}(Y))$	$\Leftrightarrow$	σ_Y
$\cos\theta$	$\Leftrightarrow$	$\rho_{X,Y}$

Note that although $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$ live in a vector space of very large dimension, the two vectors determine a two-dimensional subspace

Orthogonality Principle

- The linear estimation problem can now be recast as a geometry problem



  Find a vector $(\hat{X} - \mathrm{E}(X)) = a(Y - \mathrm{E}(Y))$ that minimizes $\|X - \hat{X}\|$

- Clearly $(X - \hat{X}) \perp (Y - \mathrm{E}(Y))$ minimizes $\|X - \hat{X}\|$, i.e.,

$$E\big((X - \hat{X})(Y - \mathrm{E}(Y)\big) \; = \; 0 \; \Rightarrow \; a \; = \; \frac{\mathrm{Cov}(X, Y)}{\sigma_Y^2}$$

- This argument is called the *orthogonality principle*. Later we will see that it is key to deriving the minimum MSE linear estimate in more complex settings

Linear vs. MMSE (Nonlinear) Estimate

- The linear estimate is not, in general, as good as the MMSE estimate
- Example: Let $Y \sim \mathrm{U}[-1, 1]$ and $X = Y^2$

  The MMSE estimate of X given Y is Y^2 — perfect!

  To find the MMSE linear estimate we compute

$$\begin{aligned} \mathrm{E}(Y) &= 0 \\ \mathrm{E}(X) &= \int_{-1}^{1} \tfrac{1}{2} y^2 \, dy = \tfrac{1}{3} \\ \mathrm{Cov}(X, Y) &= \mathrm{E}(XY) - 0 = \mathrm{E}(Y^3) = 0 \end{aligned}$$

  Thus the MMSE linear estimate $\hat{X} = \mathrm{E}(X) = \frac{1}{3}$, i.e., the observation Y is totally ignored, even though it completely determines X!

- There is a very important class of r.v.s for which the MMSE estimate is linear, the *jointly Gaussian* random variables

Jointly Guassian Random Variables

- Two r.v.s are *jointly Gaussian* if their joint pdf is of the form

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}}\, e^{-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right)}$$

- The pdf is a function only of μ_X, μ_Y, σ_X^2, σ_Y^2, and $\rho_{X,Y}$
- Note: In Lecture Notes 6 we shall define this in a more general way
- Example: For the additive Gaussian noise channel, where $X \sim \mathcal{N}(\mu, \sigma_X^2)$ and $Z \sim \mathcal{N}(0, \sigma_Z^2)$ are independent and $Y = X + Z$, (i) X and Z are jointly Gaussian, and (ii) X and Y are jointly Gaussian
- Solution: (i) It is easy to show that if two Gaussian r.v.s are independent, their joint pdf has the above form with $\rho_{X,Y} = 0$. (ii) Now consider

$$\begin{aligned} f(x,y) &= f_X(x) f_{Y|X}(y|x) \\ &= f_X(x) f_{Z|X}(y-x|x) = f_X(x) f_Z(y-x) \end{aligned}$$

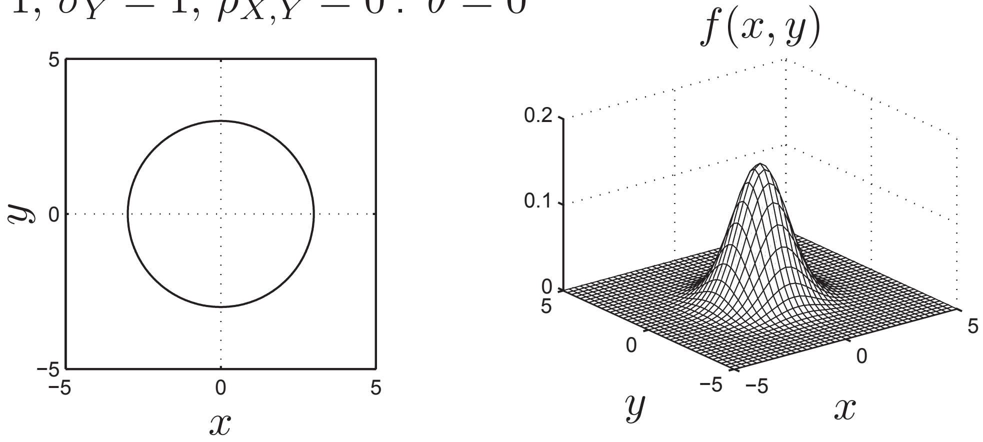
Now we can write $f(x,y)$ in the form of a jointly Gaussian pdf (here $\rho_{X,Y} > 0$)

$$\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} = c \geq 0$$

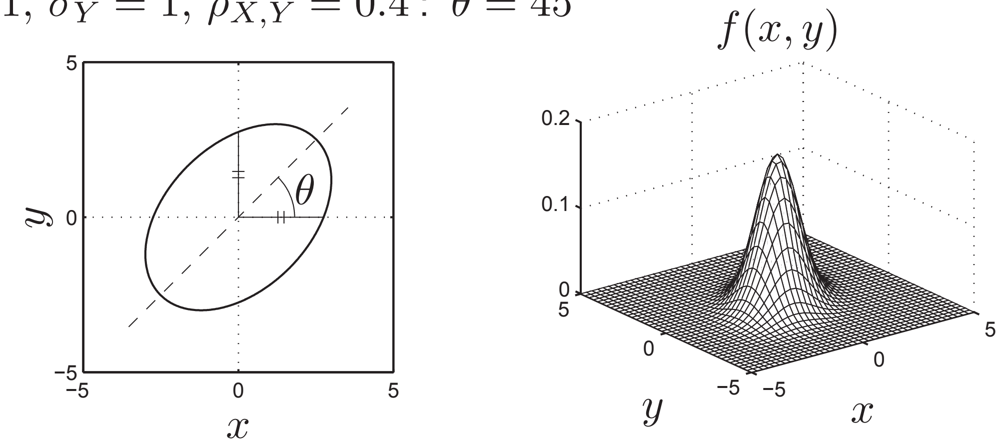
- Examples: In the following examples we plot contours of equal joint pdf $f(x,y)$ for zero mean jointly Gaussian r.v.s for different values of σ_X, σ_Y, and $\rho_{X,Y}$

  The orientation of the major axis of the ellipse is $\theta = \frac{1}{2}\arctan\left(\frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}\right)$

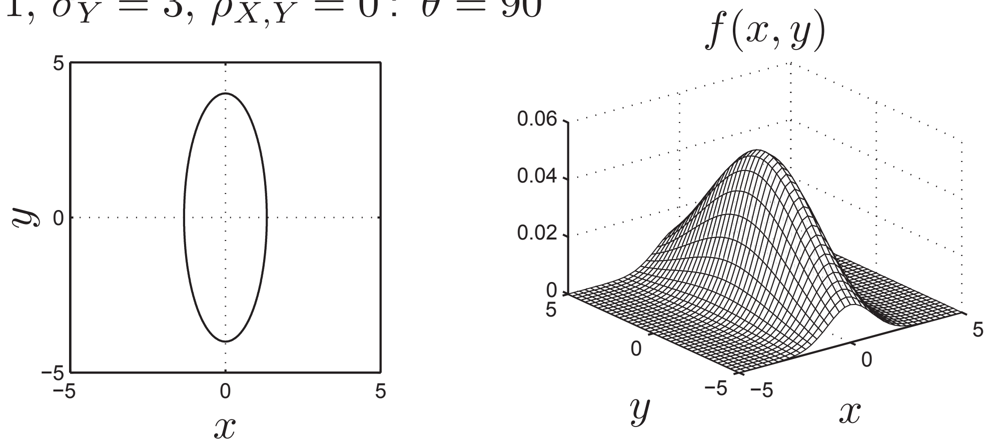
$\sigma_X = 1,\ \sigma_Y = 1,\ \rho_{X,Y} = 0:\ \theta = 0^\circ$



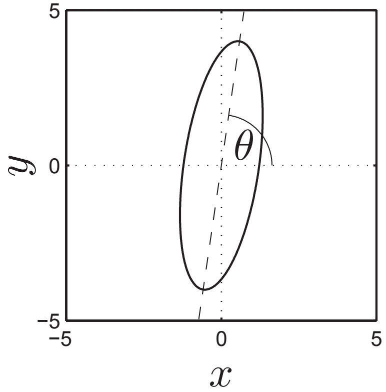
$\sigma_X = 1,\ \sigma_Y = 1,\ \rho_{X,Y} = 0.4:\ \theta = 45^\circ$

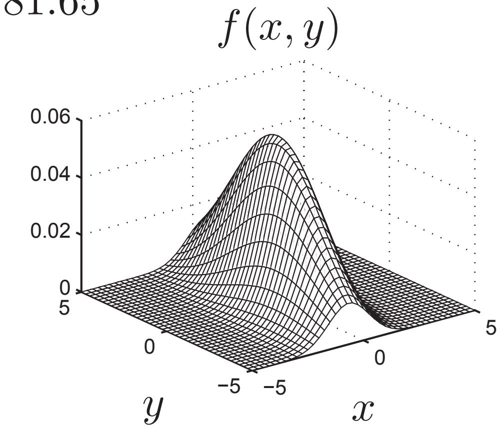


$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0:\ \theta = 90^\circ$

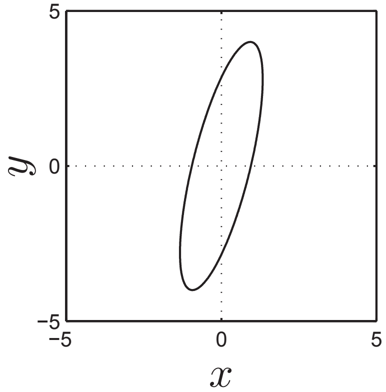


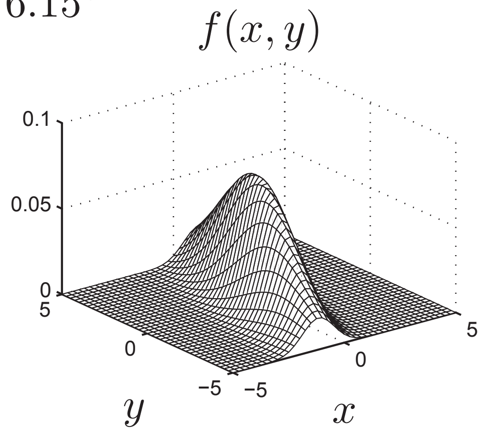
$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.4:\ \theta = 81.65^\circ$



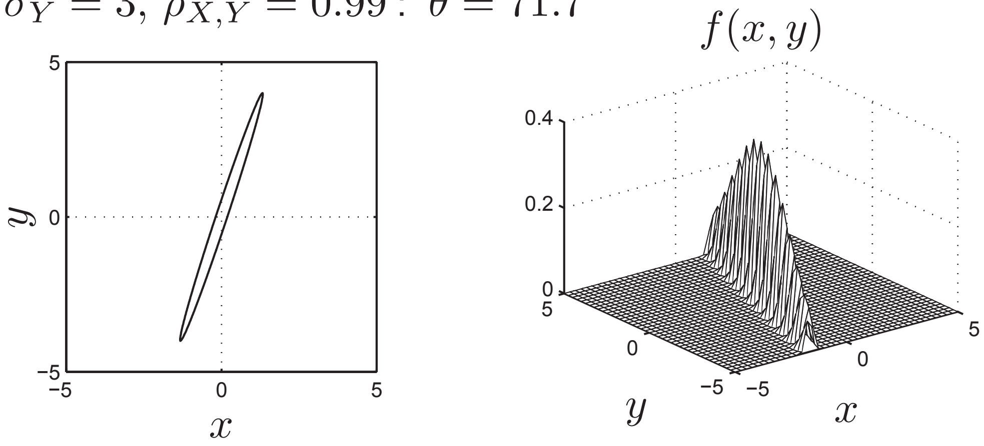


$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.7:\ \theta = 76.15^\circ$

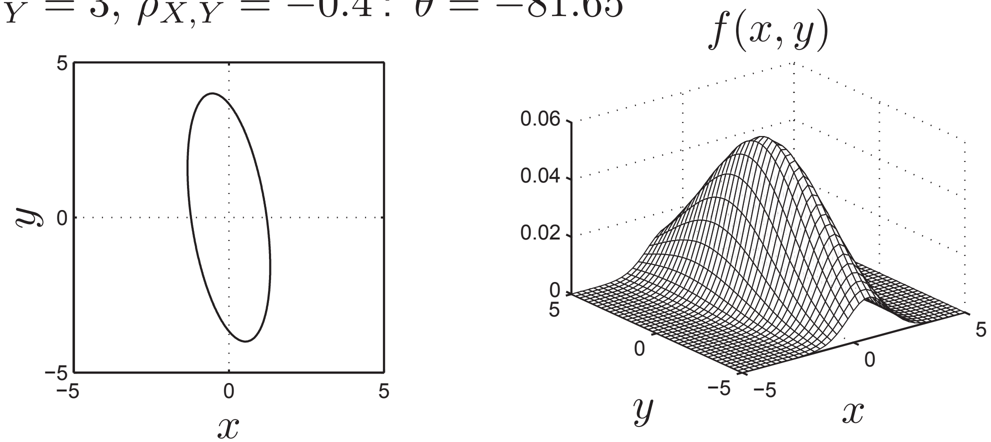




$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.99:\ \theta = 71.7°$



$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = -0.4:\ \theta = -81.65°$



Properities of Jointly Gaussian Random Variables

- If X and Y are jointly Gaussian, they are individually Gaussian, i.e., the marginals of $f_{X,Y}(x,y)$ are Gaussian, i.e.,

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)\,, \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

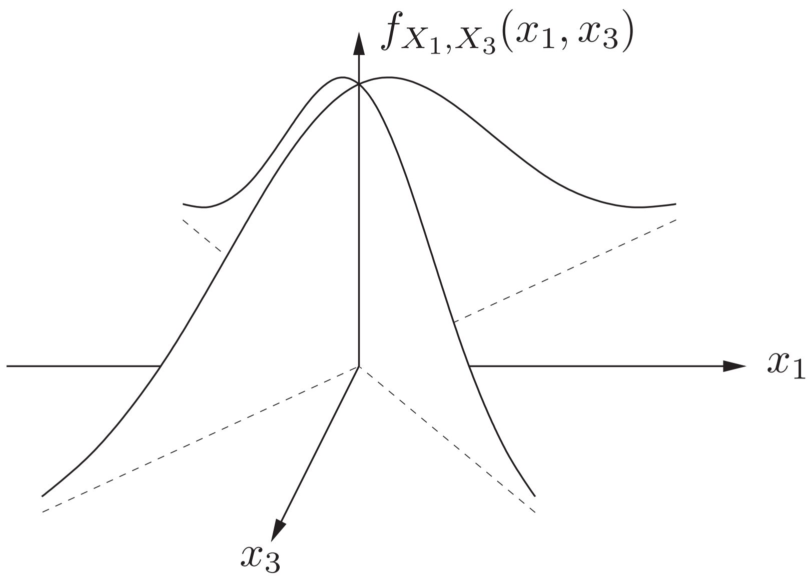
- The converse is not necessarily true, i.e., Gaussian marginals do not necessarily mean that the r.v.s are jointly Gaussian

- Example: Let $X_1 \sim \mathcal{N}(0,1)$ and

$$X_2 \;=\begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

  be independent r.v.s, and let $X_3 = X_1 X_2$

  - Clearly, $X_3 \sim \mathcal{N}(0,1)$
  - However, $f_{X_1,X_3}(x_1,x_3)$ does not have the form of the pdf of jointly Gaussian r.v.s. The pdf is shown in the following figure



- If X and Y are jointly Gaussian, the conditional pdf is Gaussian:

$$X \mid \{Y = y\} \sim \mathcal{N}\left(\rho_{X,Y}\sigma_X \frac{(y-\mu_Y)}{\sigma_Y} + \mu_X,\ (1-\rho_{X,Y}^2)\sigma_X^2\right),$$

which shows that the MMSE estimate is linear

- If X and Y are jointly Gaussian and uncorrelated, i.e., $\rho_{X,Y} = 0$, then they are also independent

Lecture 6: Random Vectors

- Joint, Marginal, and Conditional CDF, PDF, PMF
- Independence and Conditional Independence
- Mean and Covariance Matrix
- Mean and Variance of Sum of RVs
- Gaussian Random Vectors
- MSE Estimation: Vector Case

Random Vectors

- Let $X_1, X_2, \ldots, X_n$ be random variables on the same probability space. We define a *random vector* (RV) as

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

- $\mathbf{X}$ is completely specified by its joint cdf for $\mathbf{x} = (x_1, x_2, \ldots, x_n)$:

$$F_{\mathbf{X}}(\mathbf{x}) = \mathrm{P}\{X_1 \le x_1,\, X_2 \le x_2,\, \ldots,\, X_n \le x_n\}\,, \quad \mathbf{x} \in \mathbf{R}^n$$

- If $\mathbf{X}$ is continuous, i.e., $F_{\mathbf{X}}(\mathbf{x})$ is a continuous function of $\mathbf{x}$, then $\mathbf{X}$ can be specified by its joint pdf:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)\,, \quad \mathbf{x} \in \mathbf{R}^n$$

- If $\mathbf{X}$ is discrete then it can be specified by its joint pmf:

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)\,, \quad \mathbf{x} \in \mathcal{X}^n$$

- A marginal cdf (pdf, pmf) is the joint cdf (pdf, pmf) for a subset of $\{X_1, \ldots, X_n\}$; e.g., for

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

  the marginals are

$$f_{X_1}(x_1)\,,\; f_{X_2}(x_2)\,,\; f_{X_3}(x_3)$$

$$f_{X_1,X_2}(x_1, x_2)\,,\; f_{X_1,X_3}(x_1, x_3)\,,\; f_{X_2,X_3}(x_2, x_3)$$

- The marginals can be obtained from the joint in the usual way. For the previous example,

$$F_{X_1}(x_1) = \lim_{x_2, x_3 \to \infty} F_{\mathbf{X}}(x_1, x_2, x_3)$$

$$f_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1, x_2, x_3)\, dx_3$$

- Conditional cdf (pdf, pmf) can also be defined in the usual way. E.g., the conditional pdf of $\mathbf{X}_{k+1}^n = (X_{k+1}, \ldots, X_n)$ given $\mathbf{X}^k = (X_1, \ldots, X_k)$ is

$$f_{\mathbf{X}_{k+1}^n|\mathbf{X}^k}(\mathbf{x}_{k+1}^n|\mathbf{x}^k) = \frac{f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)}{f_{\mathbf{X}^k}(x_1, x_2, \ldots, x_k)} = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}^k}(\mathbf{x}^k)}$$

- *Chain Rule*: We can write

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1,X_2}(x_3|x_1, x_2) \cdots f_{X_n|\mathbf{X}^{n-1}}(x_n|\mathbf{x}^{n-1})$$

  Proof: By induction. The chain rule holds for $n = 2$ by definition of conditional pdf. Now suppose it is true for $n - 1$. Then

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{\mathbf{X}^{n-1}}(\mathbf{x}^{n-1}) f_{X_n|\mathbf{X}^{n-1}}(x_n|\mathbf{x}^{n-1}) \\ &= f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) \cdots f_{X_{n-1}|\mathbf{X}^{n-2}}(x_{n-1}|\mathbf{x}^{n-2}) f_{X_n|\mathbf{X}^{n-1}}(x_n|\mathbf{x}^{n-1}), \end{aligned}$$

  which completes the proof

Independence and Conditional Independence

- Independence is defined in the usual way; e.g., $X_1, X_2, \ldots, X_n$ are independent if

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) \quad \text{for all } (x_1, \ldots, x_n)$$

- Important special case, *i.i.d. r.v.s*: $X_1, X_2, \ldots, X_n$ are said to be *independent, identically distributed* (i.i.d.) if they are independent and have the same marginals

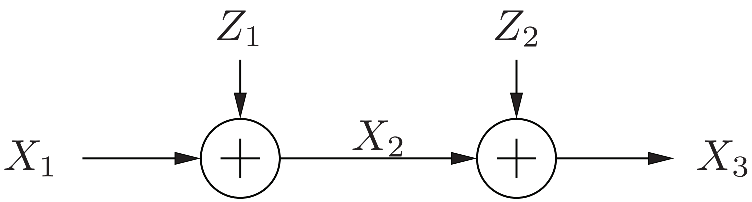
  Example: if we flip a coin n times independently, we generate i.i.d. $\mathrm{Bern}(p)$ r.v.s. $X_1, X_2, \ldots, X_n$

- R.v.s X_1 and X_3 are said to be *conditionally independent* given X_2 if

$$f_{X_1,X_3|X_2}(x_1, x_3|x_2) = f_{X_1|X_2}(x_1|x_2) f_{X_3|X_2}(x_3|x_2) \quad \text{for all } (x_1, x_2, x_3)$$

- Conditional independence neither implies nor is implied by independence; X_1 and X_3 independent given X_2 does not mean that X_1 and X_3 are independent (or vice versa)

- Example: *Serial Binary Symmetric Channel*



Here $X_1 \sim \mathrm{Bern}(p)$, $Z_1 \sim \mathrm{Bern}(\epsilon_1)$, and $Z_2 \sim \mathrm{Bern}(\epsilon_2)$, where X_1, Z_1, Z_2 are independent and $X_3 = X_1 + Z_1 + Z_2 \bmod 2 = X_1 \oplus Z_1 \oplus Z_2$

  - In general, X_1 and X_3 are not independent
  - However, X_1 and X_3 are conditionally independent given X_2
  - Also X_1 and Z_1 are independent but not conditionally independent given X_2

- Example: *Coin with Random Bias*. Given a coin with random bias $P \sim f_P(p)$, flip it n times independently to generate the r.v.s $X_1, X_2, \ldots, X_n$, where $X_i = 1$ if i-th flip is heads, 0 otherwise

  - $X_1, X_2, \ldots, X_n$ are *not* independent
  - However, $X_1, X_2, \ldots, X_n$ are conditionally independent given P; in fact, for any $P = p$, they are i.i.d. $\mathrm{Bern}(p)$

Mean and Covariance Matrix

- The mean of the random vector $\mathbf{X}$ is defined as

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix}\mathrm{E}(X_1) & \mathrm{E}(X_2) & \cdots & \mathrm{E}(X_n)\end{bmatrix}^T$$

- Denote the covariance between X_i and X_j, $\mathrm{Cov}(X_i, X_j)$, by σ_{ij} (so the variance of X_i is denoted by σ_{ii}, $\mathrm{Var}(X_i)$, or $\sigma^2_{X_i}$)

- The *covariance matrix* of $\mathbf{X}$ is defined as

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

- For $n = 2$, we can use the definition of correlation coefficient to obtain

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma^2_{X_1} & \rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} \\ \rho_{X_1,X_2}\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{bmatrix}$$

Properties of Covariance Matrix

- $\Sigma_{\mathbf{X}}$ is *real* and *symmetric* (since $\sigma_{ij} = \sigma_{ji}$)
- $\Sigma_{\mathbf{X}}$ is *nonnegative definite*, i.e., the *quadratic form*

$$\mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{a} \geq 0 \quad \text{for any real vector } \mathbf{a}$$

  Equivalently, all the *eigenvalues* of $\Sigma_{\mathbf{X}}$ are nonnegative, and also all *leading principal minors* are nonnegative

- To show that $\Sigma_{\mathbf{X}}$ is nonnegative definite we write

$$\Sigma_{\mathbf{X}} = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\right] ,$$

  i.e., as the expectation of an *outer product*. Thus

$$\begin{aligned} \mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{a} &= \mathbf{a}^T \,\mathrm{E}\left[(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\right] \mathbf{a} \\ &= \mathrm{E}\left[\mathbf{a}^T (\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T \mathbf{a}\right] \\ &= \mathrm{E}\left[(\mathbf{a}^T (\mathbf{X} - \mathrm{E}(\mathbf{X})))^2\right] \;\geq\; 0 \end{aligned}$$

Which of the following can be a Covariance Matrix

1. $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

2. $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

3. $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix}$

4. $\begin{bmatrix} -1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

5. $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}$

6. $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$

Mean and Variance of Sum of RVs

- Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a RV and let Y be the sum of the X_is. In vector notation

  $$Y = \mathbf{1}^T \mathbf{X},$$

  where $\mathbf{1}$ is the all 1 vector

  By linearity of expectation, the expected value of Y is

  $$\mathrm{E}(Y) = \mathrm{E}\left(\mathbf{1}^T \mathbf{X}\right) = \mathbf{1}^T \mathrm{E}(\mathbf{X}) = \sum_{i=1}^{n} \mathrm{E}(X_i)$$

- Example: *Mean of Binomial r.v.* One way to define a binomial r.v. is as follows: Flip a coin with bias p independently n times and define the Bernoulli r.v. X_i to be 1 if the i-th flip is a head and 0 if it is a tail. Let $Y = \sum_{i=1}^{n} X_i$. Then Y is a binomial r.v. Thus

  $$\mathrm{E}(Y) = \sum_{i=1}^{n} \mathrm{E}(X_i) = np$$

  Note that we did not need independence for this result to hold, i.e., the result holds even if the coin flips are not independent

- Let's compute the variance of Y

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{E}\left[(Y - \mathrm{E}(Y))^2\right] \\
&= \mathrm{E}\left[(\mathbf{1}^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))^2\right] \\
&= \mathrm{E}\left[\mathbf{1}^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\mathbf{1}\right] \\
&= \mathbf{1}^T\Sigma_{\mathbf{X}}\mathbf{1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))] \\
&= \sum_{i=1}^{n}\mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathrm{Cov}(X_i, X_j)
\end{aligned}
$$

If the r.v.s are independent, then $\mathrm{Cov}(X_i, X_j) = 0$, for all $i \neq j$, and

$$
\mathrm{Var}(Y) = \sum_{i=1}^{n}\mathrm{Var}(X_i)
$$

Note that this result requires only that $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, i.e., that the r.v.s are uncorrelated (which is in general weaker than independence)

- Example: *Variance of Binomial RV* Express $Y = \sum_{i=1}^{n} X_i$ where X_is are iid $\mathrm{Bern}(p)$. Since X_is are independent, $\mathrm{Cov}(X_i, X_j) = 0, \forall i \neq j$. Thus,

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p)$$

- Example: *Hats.* Suppose n people throw their hats in a box and then each picks one hat at random. Let N be the number of people that get back their own hat. Find $\mathrm{E}(N)$ and $\mathrm{Var}(N)$

  Solution: Define the r.v. $X_i = 1$ if a person selects her own hat, and $X_i = 0$, otherwise. Thus $N = \sum_{i=1}^{n} X_i$.

  To find the mean and variance of N, we first find the means, variances and covariances of the X_is

  Since $X_i \sim \mathrm{Bern}(1/n)$ we have $\mathrm{E}(X_i) = 1/n$ and $\mathrm{Var}(X_i) = (1/n)(1 - 1/n)$

  To find the covariance of X_i and X_j, $i \neq j$, note that

$$p_{X_i,X_j}(1,1) = \frac{1}{n(n-1)}$$

Thus

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\ &= \frac{1}{n(n-1)} \cdot 1 - \left(\frac{1}{n}\right)^2 = \frac{1}{n^2(n-1)}\end{aligned}$$

The mean and variance of N are given by

$$\begin{aligned}\mathrm{E}(N) &= n\,\mathrm{E}(X_1) = 1 \\ \mathrm{Var}(N) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathrm{Cov}(X_i, X_j) \\ &= n\mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1, X_2) \\ &= \left(1 - \frac{1}{n}\right) + n(n-1)\,\frac{1}{n^2(n-1)} = 1\end{aligned}$$

Method of Indicators

- In the last two examples we used the *method of indicators* to simplify the computation of expectation

- In general, the *indicator* of an event $A \subset \Omega$ is the r.v. defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

  Thus

$$\mathrm{E}[I_A] = 1 \cdot \mathrm{P}(A) + 0 \cdot \mathrm{P}(A^c) = \mathrm{P}(A)$$

- The method of indicators involves expressing a given r.v. Y as a sum of indicators in order to simplify the computation of its expectation (this is precisely what we did in the last two examples)

- Example: *Spaghetti.* We have a bowl with n spaghetti strands. You randomly pick two strand ends and join them. The process is continued until there are no ends left. Let X be the number of spaghetti loops formed. What is $\mathrm{E}(X)$?

Gaussian Random Vectors

- A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is a Gaussian random vector (GRV) (or $X_1, X_2, \ldots, X_n$ are jointly Gaussian r.v.s) if the joint pdf is of the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

  where $\boldsymbol{\mu}$ is the mean and Σ is the covariance matrix of $\mathbf{X}$, and $\left|\Sigma\right| > 0$, i.e., Σ is positive definite

- Verify that this joint pdf is the same as the case $n = 2$ from Lecture Notes 5

- Notation: $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a GRV with given mean and covariance matrix

- Since Σ is positive definite, Σ^{-1} is positive definite. Thus if $\mathbf{x} - \boldsymbol{\mu} \neq \mathbf{0}$,

$$(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) > 0,$$

  which means that the contours of equal pdf are ellipsoids

- The GRV $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, aI)$, where I is the identity matrix and $a > 0$, is called *white*; its contours of equal joint pdf are spheres centered at the origin

Properties of GRVs

- Property 1: For a GRV, uncorrelation implies independence

  This can be verified by substituting $\sigma_{ij} = 0$ for all $i \neq j$ in the joint pdf.

  Then Σ becomes diagonal and so does Σ^{-1}, and the joint pdf reduces to the product of the marginals $X_i \sim \mathcal{N}(\mu_i, \sigma_{ii})$

  For the white GRV $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, aI)$, the r.v.s are i.i.d. $\mathcal{N}(0, a)$

- Property 2: Linear transformation of a GRV yields a GRV, i.e., given any $m \times n$ matrix A, where $m \leq n$ and A has full rank m, then

$$\mathbf{Y} = A\mathbf{X} \sim \mathcal{N}(A\boldsymbol{\mu},\ A\Sigma A^T)$$

- Example: Let

$$\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\right)$$

  Find the joint pdf of

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{X}$$

- Solution: From Property 2, we have

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix}\right)$$

Before we prove Property 2, let us show that

$$\mathrm{E}(\mathbf{Y}) = A\boldsymbol{\mu} \quad \text{and} \quad \Sigma_{\mathbf{Y}} = A\Sigma A^T$$

These results follow from linearity of expectation. First, expectation:

$$\mathrm{E}(\mathbf{Y}) = \mathrm{E}(A\mathbf{X}) = A\,\mathrm{E}(\mathbf{X}) = A\boldsymbol{\mu}$$

Next consider the covariance matrix:

$$\begin{aligned} \Sigma_{\mathbf{Y}} &= \mathrm{E}\left[(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))^T\right] \\ &= \mathrm{E}\left[(A\mathbf{X} - A\boldsymbol{\mu})(A\mathbf{X} - A\boldsymbol{\mu})^T\right] \\ &= A\,\mathrm{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right] A^T = A\Sigma A^T \end{aligned}$$

Of course this is not sufficient to show that $\mathbf{Y}$ is a GRV—we must also show that the joint pdf has the right form

We do so using the *characteristic function* for a random vector

- Definition: If $\mathbf{X} \sim f_{\mathbf{X}}(\mathrm{X})$, the characteristic function of $\mathbf{X}$ is $\Phi_{\mathbf{X}}(\omega) = \mathrm{E}\left(e^{i\omega^{\top}\mathbf{X}}\right)$,

  where $\boldsymbol{\omega}$ is an n-dimensional real valued vector and $i = \sqrt{-1}$

  Thus

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) e^{i\boldsymbol{\omega}^T\mathbf{x}}\, d\mathbf{x}$$

  This is the inverse of the multi-dimensional Fourier transform of $f_{\mathbf{X}}(\mathbf{x})$, which implies that there is a one-to-one correspondence between $\Phi_{\mathbf{X}}(\boldsymbol{\omega})$ and $f_{\mathbf{X}}(\mathbf{x})$, which can be found by taking the Fourier transform

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{1}{(2\pi)^n} \Phi_{\mathbf{X}}(\boldsymbol{\omega}) e^{-i\boldsymbol{\omega}^T\mathbf{x}}\, d\boldsymbol{\omega}$$

- Example: The characteristic function for $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$\Phi_X(\omega) = e^{-\frac{1}{2}\omega^2\sigma^2 + i\mu\omega},$$

  and for a GRV $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$,

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = e^{-\frac{1}{2}\boldsymbol{\omega}^T\Sigma\boldsymbol{\omega} + i\boldsymbol{\omega}^T\boldsymbol{\mu}}$$

- Now let's go back to proving Property 2

  Since A is an $m \times n$ matrix, $\mathbf{Y} = A\mathbf{X}$ and $\boldsymbol{\omega}$ are m-dimensional. Therefore the characteristic function of $\mathbf{Y}$ is

$$\begin{aligned}
\Phi_{\mathbf{Y}}(\boldsymbol{\omega}) &= \mathrm{E}\left(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\right) \\
&= \mathrm{E}\left(e^{i\boldsymbol{\omega}^T A\mathbf{X}}\right) \\
&= \Phi_{\mathbf{X}}(A^T\boldsymbol{\omega}) \\
&= e^{-\frac{1}{2}(A^T\boldsymbol{\omega})^T\Sigma(A^T\boldsymbol{\omega}) + i\boldsymbol{\omega}^T A\boldsymbol{\mu}} \\
&= e^{-\frac{1}{2}\boldsymbol{\omega}^T(A\Sigma A^T)\boldsymbol{\omega} + i\boldsymbol{\omega}^T A\boldsymbol{\mu}}
\end{aligned}$$

  Thus $\mathbf{Y} = A\mathbf{X} \sim \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^T)$

- An equivalent definition of GRV: $\mathbf{X}$ is a GRV iff for any real vector $\mathbf{a} \neq 0$, the r.v. $Y = \mathbf{a}^T\mathbf{X}$ is Gaussian (see HW for proof)

- Property 3: Marginals of a GRV are Gaussian, i.e., if $\mathbf{X}$ is GRV, for any subset $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ of indices, the RV

$$\mathbf{Y} = \begin{bmatrix} X_{i_1} \\ X_{i_2} \\ \vdots \\ X_{i_k} \end{bmatrix}$$

  is a GRV

  - To show this we use Property 2. For example, let $n = 3$ and $\mathbf{Y} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$

    We can express $\mathbf{Y}$ as a linear transformation of $\mathbf{X}$:

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$$

    Therefore

$$\mathbf{Y} \sim \mathcal{N}\left(\begin{bmatrix} \mu_3 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \sigma_{33} & \sigma_{31} \\ \sigma_{13} & \sigma_{11} \end{bmatrix} \right)$$

  - The converse of Property 3 does not hold in general (as demonstrated by the example in Lecture Notes 5)

- Property 4: Conditionals of a GRV are Gaussian, more specifically, if

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ -- \\ \mathbf{X}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ -- \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & | & \Sigma_{12} \\ -- & | & -- \\ \Sigma_{21} & | & \Sigma_{22} \end{bmatrix} \right),$$

  where $\mathbf{X}_1$ is a k-dim RV and $\mathbf{X}_2$ is an $n-k$-dim RV, then

$$\mathbf{X}_2 \,|\, \{\mathbf{X}_1 = \mathbf{x}\} \sim \mathcal{N}\left(\Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2\,,\ \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

  Compare this to the case of $n = 2$ and $k = 1$:

$$X_2 \,|\, \{X_1 = x\} \sim \mathcal{N}\left(\frac{\sigma_{21}}{\sigma_{11}}(x - \mu_1) + \mu_2\,,\ \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)$$

- Example:

$$\begin{bmatrix} X_1 \\ -- \\ X_2 \\ X_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 1 \\ -- \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & | & 2 & 1 \\ -- & | & -- & -- \\ 2 & | & 5 & 2 \\ 1 & | & 2 & 9 \end{bmatrix} \right)$$

From Property 4, it follows that

$$\mathrm{E}(\mathbf{X}_2 \,|\, X_1 = x) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} (x-1) + \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2x \\ x+1 \end{bmatrix}$$

$$\begin{aligned} \Sigma_{\{\mathbf{X}_2 | X_1 = x\}} &= \begin{bmatrix} 5 & 2 \\ 2 & 9 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix} \end{aligned}$$

- The proof of Property 4 follows from properties 1 and 2 and the orthogonality principle (HW exercise)

- A consequence of Property 4 is that if $\left[\mathbf{Y}^T \; X\right]^T$ is a GRV, then the best MSE estimate of X given $\mathbf{Y}$ is linear, i.e., the linear MMSE estimate is the MMSE estimate

MSE Estimation: Vector Case

- Let $X \sim f_X(x)$ be a r.v. representing the signal and let $\mathbf{Y}$ be an n-dimensional RV representing the observations
- The minimum MSE estimate of X given $\mathbf{Y}$ is the conditional expectation $\mathrm{E}(X \mid \mathbf{Y})$. This is often not practical to compute either because the conditional pdf of X given $\mathbf{Y}$ is not known or because of high computational cost
- The MMSE linear (or affine) estimate is easier to find since it depends only on the means, variances, and covariances of the r.v.s involved
- To find the MMSE linear estimate, first assume that $\mathrm{E}(X) = 0$ and $\mathrm{E}(\mathbf{Y}) = \mathbf{0}$. The problem reduces to finding a real n-vector $\mathbf{h}$ such that

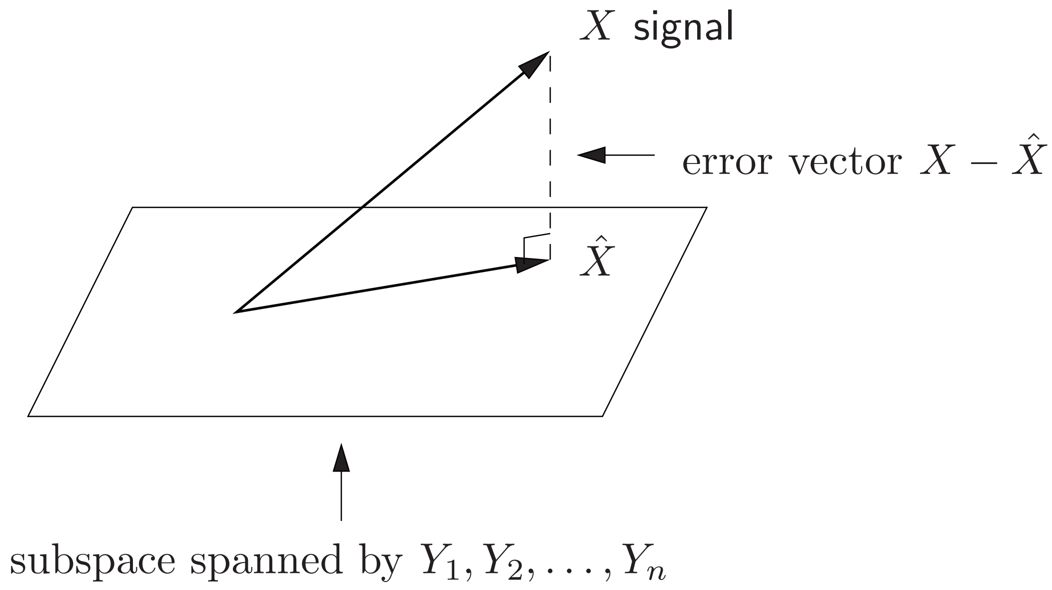
$$\hat{X} = \mathbf{h}^T \mathbf{Y} = \sum_{i=1}^{n} h_i Y_i$$

  minimizes the $\mathrm{MSE} = \mathrm{E}\left[(X - \hat{X})^2\right]$

MMSE Linear Estimate via Orthogonality Principle

- To find $\hat{X}$ we use the orthogonality principle: we view the r.v.s $X, Y_1, Y_2, \ldots, Y_n$ as vectors in the inner product space consisting of all zero mean r.v.s defined over the underlying probability space

- The linear estimation problem reduces to a geometry problem



- To minimize $\mathrm{MSE} = ||X - \hat{X}||^2$, $\hat{X}$ is chosen such that $X - \hat{X}$ is orthogonal to the subspace spanned by the observations $Y_1, \ldots, Y_n$, i.e.,

$$\mathrm{E}\left[(X - \hat{X})Y_i\right] = 0\,, \quad i = 1, 2, \ldots, n\,,$$

  hence

$$\mathrm{E}(Y_i X) = \mathrm{E}(Y_i \hat{X}) = \sum_{j=1}^{n} h_j \, \mathrm{E}(Y_i Y_j)\,, \quad i = 1, 2, \ldots, n$$

- Define the *cross covariance* of $\mathbf{Y}$ and X as the n-vector

$$\Sigma_{\mathbf{Y}X} = \mathrm{E}\left[(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))(X - \mathrm{E}(X))\right] = \begin{bmatrix} \sigma_{Y_1 X} \\ \sigma_{Y_2 X} \\ \vdots \\ \sigma_{Y_n X} \end{bmatrix}$$

  For $n = 1$ this is simply the covariance

- The above equations can be written in vector form as $\Sigma_{\mathbf{Y}} \mathbf{h} = \Sigma_{\mathbf{Y}X}$

- If $\Sigma_{\mathbf{Y}}$ is nonsingular, we can solve the equations to obtain $\mathbf{h} = \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X}$

- Thus, if $\Sigma_{\mathbf{Y}}$ is not singular, the best linear MSE estimate is $\Sigma_{\mathbf{Y}X}^{\top}\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y}$.

  - Now to find the minimum MSE, consider

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}\left[(X-\hat{X})^2\right] \\
&= \mathrm{E}\left[(X-\hat{X})X\right] - \mathrm{E}\left[(X-\hat{X})\hat{X}\right] \\
&= \mathrm{E}\left[(X-\hat{X})X\right], \text{ since by orthogonality } (X-\hat{X}) \perp \hat{X} \\
&= \mathrm{E}(X^2) - \mathrm{E}(\hat{X}X) \\
&= \sigma_X^2 - \mathrm{E}\left(\Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y}X\right) = \sigma_X^2 - \Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}X}
\end{aligned}$$

  - Compare this to the scalar case, where minimum MSE is $\sigma_X^2 - \dfrac{\mathrm{Cov}(X,Y)^2}{\sigma_Y^2}$

  - If X or $\mathbf{Y}$ have nonzero mean, the MMSE affine estimate $\hat{X} = h_0 + \mathbf{h}^T\mathbf{Y}$ is determined by first finding the MMSE linear estimate of $X - \mathrm{E}(X)$ given $\mathbf{Y} - \mathrm{E}(\mathbf{Y})$ (minimum MSE for $\hat{X}'$ and $\hat{X}$ are the same), which is $\hat{X}' = \Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))$, and then setting $\hat{X} = \hat{X}' + \mathrm{E}(X)$ (since $\mathrm{E}(\hat{X}) = \mathrm{E}(X)$ is necessary)

Example

- Let X be the r.v. representing a signal with mean μ and variance P. The observations are $Y_i = X + Z_i$, for $i = 1, 2, \ldots, n$, where the Z_i are zero mean uncorrelated noise with variance N, and X and Z_i are also uncorrelated

  Find the MMSE linear estimate of X given $\mathbf{Y}$ and its MSE

- For $n = 1$, we already know that $\hat{X}_1 = \dfrac{P}{P+N} Y_1 + \dfrac{N}{P+N}\mu$

- To find the MMSE linear estimate for general n, first let $X' = X - \mu$ and $Y_i' = Y_i - \mu$. Thus X' and $\mathbf{Y}'$ are zero mean

- The MMSE linear estimate of X' given $\mathbf{Y}'$ is given by $\hat{X}_n' = \mathbf{h}^T \mathbf{Y}'$, where

$$\Sigma_{\mathbf{Y}} \mathbf{h} = \Sigma_{\mathbf{Y}X}, \quad \text{thus}$$

$$\begin{bmatrix} P+N & P & \cdots & P \\ P & P+N & \cdots & P \\ \vdots & \vdots & \ddots & \vdots \\ P & P & \cdots & P+N \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} = \begin{bmatrix} P \\ P \\ \vdots \\ P \end{bmatrix}$$

- By symmetry, $h_1 = h_2 = \cdots = h_n = \dfrac{P}{nP + N}$. Thus

$$\hat{X}_n' = \frac{P}{nP+N} \sum_{i=1}^{n} Y_i'$$

Therefore

$$\hat{X}_n = \frac{P}{nP+N}\left(\sum_{i=1}^{n}(Y_i - \mu)\right) + \mu = \frac{P}{nP+N}\left(\sum_{i=1}^{n} Y_i\right) + \frac{N}{nP+N}\mu$$

- The mean square error of the estimate:

$$\mathrm{MSE}_n = P - \mathrm{E}(\hat{X}_n' X') = \frac{PN}{nP+N}$$

Thus as $n \to \infty$, $\mathrm{MSE}_n \to 0$, i.e., the linear estimate becomes perfect (even though we don't know the complete statistics of X and Y)