5.1 DEFINITION AND PROPERTIES

Let $X \in \mathcal{X}$ be a discrete random variable with pmf $p_X(x)$ and let $g(x)$ be a function of $x$. The expectation (or expected value or mean) of $g(X)$ is defined as

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)p_X(x).$$

For a continuous random variable $X \sim f_X(x)$, the expected value of $g(X)$ is defined as

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x) \, dx.$$

The expectation operation $\mathbb{E}[\cdot]$ satisfies the following properties:

1. $\mathbb{E}[c] = c$ for every constant $c$.
2. If $g(X) \geq 0$ w.p. 1, then $\mathbb{E}[g(X)] \geq 0$.
3. Linearity. For any constant $a$ and functions $g_1(x)$ and $g_2(x)$,

$$\mathbb{E}[ag_1(X) + g_2(X)] = a\mathbb{E}[g_1(X)] + \mathbb{E}[g_2(X)].$$

By considering $Y = g(X)$ as a random variable on its own, we can compute the same expectation.

**Fundamental theorem of expectation.** If $X \sim p_X(X)$ and $Y = g(X) \sim p_Y(y)$, then

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} y p_Y(y) = \sum_{x \in \mathcal{X}} g(x)p_X(x) = \mathbb{E}[g(X)].$$

Similarly, if $X \sim f_X(x)$ and $Y = g(X) \sim f_Y(y)$, then

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) \, dy = \int_{-\infty}^{\infty} g(x)f_X(x) \, dx = \mathbb{E}[g(X)].$$
We prove the theorem for the discrete case. Consider
\[
E[Y] = \sum_y y p_Y(y)
\]
\[
= \sum_y \sum_{x: g(x)=y} p_X(x)
\]
\[
= \sum_y \sum_{x: g(x)=y} y p_X(x)
\]
\[
= \sum_y \sum_{x: g(x)=y} g(x) p_X(x)
\]
\[
= \sum_x g(x) p_X(x).
\]

Thus, \(E[Y] = E[g(X)]\) can be found using either \(p_X(x)\) or \(p_Y(y)\). It is often much easier to use \(p_X(x)\) than to first find \(p_Y(y)\) and then find \(E[Y]\).

We already know that a random variable is completely specified, that is, any probability of a Borel set involving the random variable can be determined, by its cumulative distribution function (or its pmf and pdf in discrete and continuous cases, respectively). As a simple summary of the random variable, however, its expectation has several applications.

1. Expectation can be used to bound or estimate probabilities of interesting events, as we will see in Section 5.3.
2. Expectation provides the optimal estimate of a random variable under the mean square error criterion, as we will see in Section 5.5.
3. It is far easier to estimate the expectation of a random variable from data than to estimate its distribution, as we will see in Lecture #7.

### 5.2 MEAN AND VARIANCE

The first moment (or mean) of \(X \sim f_X(x)\) is the expectation
\[
E[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx.
\]

The following trick is useful for computing the expectation of a nonnegative random variable. If \(X \geq 0\) is continuous, then
\[
E[X] = \int_0^\infty u f_X(u) \, du
\]
\[
= \int_0^\infty \int_0^u dx f_X(u) \, du
\]
\[
= \int_0^\infty \int_x^\infty f_X(u) \, du \, dx
\]
\[
= \int_0^\infty 1 - F_X(x) \, dx.
\]
The same identity also follows by integration by parts \( \int \phi' \psi = \phi' \psi - \int \phi' \psi \) with \( \phi = x \) and \( \psi = 1 - F_X(x) \). Similarly, if \( X \geq 0 \) is discrete, then
\[
E[X] = \sum_{k=0}^{\infty} 1 - F_X(k).
\]

Let
\[
1_A(x) = \begin{cases} 
1 & x \in A, \\
0 & \text{otherwise}.
\end{cases}
\]

Then, the expectation of the indicator variable is
\[
E[1_A(X)] = \int_{-\infty}^{\infty} 1_A(x) f_X(x) \, dx = \int_{A} f_X(x) \, dx = P\{X \in A\}.
\]

The second moment (or mean square or average power) of \( X \) is
\[
E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx.
\]

The variance of \( X \) is
\[
\text{Var}(X) = E[(X - E(X))^2] \\
= E[X^2 - 2X E(X) + (E(X))^2] \\
= E[X^2] - 2(E[X])^2 + (E[X])^2 \\
= E[X^2] - (E[X])^2.
\]

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

The following useful identities can be proved by linearity of expectation:
\[
E(aX + b) = a E(X) + b \\
\text{Var}(aX + b) = a^2 \text{Var}(X)
\]

Table 5.1 summarizes the mean and variance of famous random variables.

**Remark 5.1.** Expectation can be infinite. For example, consider
\[
f_X(x) = \begin{cases} 
1/x^2 & 1 \leq x < \infty \\
0 & \text{otherwise}.
\end{cases}
\]

Then
\[
E(X) = \int_{1}^{\infty} x/x^2 \, dx = \infty.
\]
### Table 5.1. The mean and variance of common random variables.

<table>
<thead>
<tr>
<th>Random Variable</th>
<th>Mean $\mu$</th>
<th>Variance $\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Bern}(p)$</td>
<td>$p$</td>
<td>$p(1-p)$</td>
</tr>
<tr>
<td>$\text{Geom}(p)$</td>
<td>$\frac{1}{p}$</td>
<td>$\frac{1-p}{p^2}$</td>
</tr>
<tr>
<td>$\text{Binom}(n, p)$</td>
<td>$np$</td>
<td>$np(1-p)$</td>
</tr>
<tr>
<td>$\text{Poisson}(\lambda)$</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>$\text{Unif}[a, b]$</td>
<td>$\frac{a+b}{2}$</td>
<td>$\frac{(b-a)^2}{12}$</td>
</tr>
<tr>
<td>$\text{Exp}(\lambda)$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda^2}$</td>
</tr>
<tr>
<td>$\mathcal{N}(\mu, \sigma^2)$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

**Remark 5.2.** Expectation may not exist. To find conditions under which expectation exists, consider

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

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

**Example 5.1.** 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 $E(X^2) = \infty$, so it exists.)

### 5.3 Inequalities

In many cases we do not know the distribution of a random variable $X$, but wish to find the probability of an event such as $\{X > a\}$ or $\{|X - E(X)| > a\}$. The Markov and Chebyshev inequalities provide upper bounds on the probabilities of such events in terms of the mean and variance of the random variable.

**Markov inequality.** Let $X \geq 0$ be a random variable with finite mean. Then for any $a > 1$,

$$P\{X \geq a E[X]\} \leq \frac{1}{a}.$$
Example 5.2. If the average age in the San Diego is 35, then at most half of the population is 70 or older.

To prove the Markov inequality, let \( A = \{ x \geq a \, \mathbb{E}(X) \} \) and consider the indicator function \( 1_A(x) \). As illustrated in Figure 5.1.

\[
1_A(x) \leq \frac{x}{a \, \mathbb{E}[X]}.
\]

Since \( \mathbb{E}(1_A(X)) = \mathbb{P}(X \geq a \, \mathbb{E}[X]) \), taking the expectations of both sides establishes the inequality.

The Markov inequality can be very loose. For example, if \( X \sim \text{Exp}(1) \), then

\[
\mathbb{P}(X \geq 10) = e^{-10} \approx 4.54 \times 10^{-5}.
\]

The Markov inequality yields

\[
\mathbb{P}(X \geq 10) \leq \frac{1}{10},
\]

which is very pessimistic. But it is the tightest possible inequality on \( \mathbb{P}(X \geq a \, \mathbb{E}[X]) \) when we are given only \( \mathbb{E}[X] \). To show this, note that the inequality is tight for

\[
X = \begin{cases} 
\mathbb{E}[X] & \text{w.p. } 1/a, \\
0 & \text{w.p. } 1 - 1/a.
\end{cases}
\]

In Example 5.2 if half of the population is 0 year old and half other population is 70 years old, then the average age is 35 and the Markov inequality is tight.

**Chebyshev inequality.** Let \( X \) be a random variable with finite mean \( \mathbb{E}[X] \) and variance \( \sigma_X^2 \). Then for any \( a > 1 \),

\[
\mathbb{P}(|X - \mathbb{E}[X]| \geq a\sigma_X) \leq \frac{1}{a^2}.
\]
Example 5.3. 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. Then, by the Chebyshev inequality, the fraction of out-of-spec ICs, namely, $P(|X - E(X)| \geq 3\sigma_X)$ is no larger than $\frac{1}{9}$.

The proof of the Chebyshev inequality uses the Markov inequality (which is a slight twist from the teacher–student relationship between Pafnuty Chebyshev and Andrey Markov at Saint Petersburg University in Russia). Define the random variable $Y = (X - E[X])^2 \geq 0$. Since $E[Y] = \sigma_X^2$, the Markov inequality implies that

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

The Chebyshev inequality can be very loose. Let $X \sim N(0, 1)$. Then, by the Chebyshev inequality,

$$P(|X| \geq 3) \leq \frac{1}{7},$$

which is very pessimistic compared to the actual value $2Q(3) \approx 2 \times 10^{-3}$. But it is the tightest upper bound on $P(|X - E(X)| \geq a\sigma_X)$ given knowledge only of the mean and variance of $X$. Indeed, the inequality holds with equality for the random variable

$$X = \begin{cases} E(X) + a\sigma_X & \text{w.p. } 1/2a^2, \\ E(X) - a\sigma_X & \text{w.p. } 1/2a^2, \\ E(X) & \text{w.p. } 1 - 1/a^2. \end{cases}$$

We now discuss an extremely useful inequality that is named after a Danish mathematician Johan Jensen and is centered around the notion of convexity. A function $g(x)$ is said to be convex if

$$g(x) \leq \frac{g(b) - g(a)}{b - a}(x - a) + g(a)$$

for all $x \in [a, b]$ and all $a < b$. If $g(x)$ is twice differentiable, then $g(x)$ is convex iff

$$g''(x) \geq 0.$$ 

If $-g(x)$ is convex, then $g(x)$ is called concave.

Example 5.4. The following functions are convex: (a) $g(x) = ax + b$. (b) $g(x) = x^2$. (c) $g(x) = |x|^p, p \geq 1$. (d) $g(x) = x \log x, x > 0$. (e) $g(x) = 1/x, x > 0$.

Example 5.5. The following functions are concave: (a) $g(x) = ax + b$. (b) $g(x) = \sqrt{x}, x > 0$. (c) $g(x) = \log x, x > 0$. 
**Jensen’s inequality.** Let $X$ be a random variable with finite mean $E[X]$ and $g(x)$ be a function such that $E[g(X)]$ is finite. If $g(x)$ is convex, then

$$E[g(X)] \geq g(E[X]).$$

If $g(x)$ is concave, then

$$E[g(X)] \leq g(E[X]).$$

Jensen’s inequality is a powerful tool to derive many inequalities.

**Example 5.6.** Since $g(x) = x^2$ is convex,

$$E[g(X)] = E[X^2] \geq g(E[X]) = (E[X])^2,$$

which we already know since $\text{Var}(X) = E[X^2] - (E[X])^2 \geq 0$.

**Example 5.7.** Since $g(x) = 1/x$, $x > 0$, is convex,

$$E\left[\frac{1}{X^2}\right] \leq \frac{1}{E[X^2]}.$$

**Example 5.8 (Monotonicity of norms).** For $1 \leq p \leq q$,

$$(E[|X|^p])^{1/p} \leq (E[|X|^q])^{1/q}. \quad (5.1)$$

To see this, consider a convex function $g(x) = |x|^{q/p}$ and use Jensen’s inequality to obtain

$$E[|X|^q] = E[(|X|^p)^{q/p}] \geq (E[|X|^p])^{q/p}.$$

Taking the $q$-th root of both sides establishes (5.1).

### 5.4 Covariance and Correlation

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

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

As an example, the correlation of $X$ and $Y$ is defined as

$$E[XY].$$

We say that $X$ and $Y$ are orthogonal if $E(XY) = 0$. 

---

**Example 5.6.** Since $g(x) = x^2$ is convex,

$$E[g(X)] = E[X^2] \geq g(E[X]) = (E[X])^2,$$

which we already know since $\text{Var}(X) = E[X^2] - (E[X])^2 \geq 0$.

**Example 5.7.** Since $g(x) = 1/x$, $x > 0$, is convex,

$$E\left[\frac{1}{X^2}\right] \leq \frac{1}{E[X^2]}.$$

**Example 5.8 (Monotonicity of norms).** For $1 \leq p \leq q$,

$$(E[|X|^p])^{1/p} \leq (E[|X|^q])^{1/q}. \quad (5.1)$$

To see this, consider a convex function $g(x) = |x|^{q/p}$ and use Jensen’s inequality to obtain

$$E[|X|^q] = E[(|X|^p)^{q/p}] \geq (E[|X|^p])^{q/p}.$$

Taking the $q$-th root of both sides establishes (5.1).
The covariance of $X$ and $Y$ is defined as
\[
\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]
\]
\[
\]
\[
= E[XY] - E[X] E[Y].
\]

We say that $X$ and $Y$ are uncorrelated if $\text{Cov}(X, Y) = 0$. Note that $\text{Cov}(X, X) = \text{Var}(X)$.

The correlation coefficient of $X$ and $Y$ is defined as
\[
\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.
\]

For any pair of random variables $X$ and $Y$,
\[
|\rho_{X,Y}| \leq 1,
\]
which follows by the Cauchy–Schwarz inequality
\[
(E[XY])^2 \leq E[X^2] E[Y^2].
\]

Note that $\rho_{X,Y} = \pm 1$ iff
\[
\frac{X - E[X]}{\sigma_X} = \pm \frac{Y - E[Y]}{\sigma_Y},
\]
that is, iff $X - E[X]$ is a linear function of $Y - E[Y]$. We shall see in Section 5.6 that $\rho_{X,Y}$ is a measure of how closely $X - E[X]$ can be approximated or estimated by a linear function of $Y - E[Y]$.

**Example 5.9.** We find $E(X)$, $\text{Var}(X)$, and $\text{Cov}(X, Y)$ for $(X, Y) \sim f(x, y)$ where
\[
f(x, y) = \begin{cases} 2 & x \geq 0, \ y \geq 0, \ x + y \leq 1, \\ 0 & \text{otherwise}. \end{cases}
\]

Consider
\[
E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) \, dx \, dy
\]
\[
= \int_0^1 \int_0^{1-x} 2x \, dy \, dx
\]
\[
= 2 \int_0^1 (1-x)x \, dx
\]
\[
= 2 \left( \frac{1}{2} - \frac{1}{3} \right) = \frac{1}{3}.
\]

Since
\[
E[X^2] = 2 \int_0^1 (1-x)x^2 \, dx = 2 \left( \frac{1}{3} - \frac{1}{4} \right) = \frac{1}{6},
\]
we have
\[ \text{Var}(X) = \text{E}[X^2] - (\text{E}[X])^2 = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}. \]

By symmetry, \( \text{E}[Y] = \text{E}[X] = \frac{1}{3}. \) Thus the covariance of \( X \) and \( Y \) is
\[
\text{Cov}(X, Y) = 2 \int_0^1 \int_0^{1-x} xy \, dy \, dx - \text{E}[X] \text{E}[Y]
\]
\[
= \int_0^1 x(1-x)^2 \, dx - \frac{1}{9}
\]
\[
= \frac{1}{12} - \frac{1}{9} = -\frac{1}{36}.
\]

As noted earlier, \( X \) and \( Y \) are **uncorrelated** if \( \text{Cov}(X, Y) = 0. \) If \( X \) and \( Y \) are independent, then they are uncorrelated, since
\[
\text{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x, y) \, dx \, dy
\]
\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) \, dx \, dy
\]
\[
= \left( \int_{-\infty}^{\infty} x f_X(x) \, dx \right) \left( \int_{-\infty}^{\infty} y f_Y(y) \, dy \right)
\]
\[
= \text{E}[X] \text{E}[Y].
\]

However, that \( X \) and \( Y \) are uncorrelated does **not** necessarily imply that they are independent.

**Example 5.10.** Consider the pmf \( p_{X,Y}(x, y) \) described by the following table

<table>
<thead>
<tr>
<th>( x )</th>
<th>-1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>( \frac{1}{6} )</td>
<td>0</td>
<td>( \frac{1}{6} )</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>( \frac{1}{5} )</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>( \frac{1}{6} )</td>
<td>0</td>
<td>( \frac{1}{6} )</td>
</tr>
</tbody>
</table>

Clearly \( X \) and \( Y \) are not independent. But it can be readily checked that \( \text{E}[X] = \text{E}[Y] = \text{E}[XY] = 0. \) Thus \( \text{Cov}(X, Y) = 0, \) that is, \( X \) and \( Y \) are uncorrelated.

### 5.5 CONDITIONAL EXPECTATION

Let \((X, Y) \sim f_{X,Y}(x, y)\). Recall that the **conditional pdf** of \( X \) given \( Y = y \) is
\[
f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},
\]
if \( f_Y(y) > 0 \). Since \( f_{X|Y}(x|y) \) is a pdf for \( X \) (for each \( y \)), we can define the expectation of any function \( g(X, Y) \) w.r.t. \( f_{X|Y}(x|y) \) as

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

which is a function of \( y \).

**Example 5.11.** If \( g(X, Y) = X \), then the conditional expectation of \( X \) given \( Y = y \) is

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

**Example 5.12.** If \( g(X, Y) = Y \), then \( E[Y \mid Y = y] = y \).

**Example 5.13.** If \( g(X, Y) = XY \), then \( E[XY \mid Y = y] = y E[X \mid Y = y] \).

**Example 5.14.** Let

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

From Lecture #4, we already know that \( X \mid \{Y = y\} \sim \text{Unif}[0, 1 - y] \). Thus, \( E[X \mid Y = y] = (1 - y)/2 \).

Let \( \phi(y) = E[g(X, Y) \mid Y = y] \). We define the *conditional expectation* of \( g(X, Y) \) given \( Y \) as

\[
E[g(X, Y) \mid Y] = \phi(Y).
\]

In other words, the random variable \( E[g(X, Y) \mid Y] \) is a function of \( Y \) that takes values \( E[g(X, Y) \mid Y = y] \).

**Law of iterated expectation.** The following observation is very useful in computing expectation:

\[
E[E[g(X, Y) \mid Y]] = \int_{-\infty}^{\infty} E[g(X, Y) \mid Y = y] f_Y(y) \, dy
\]

\[
= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) \, dx \right) f_Y(y) \, dy
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) \, dx \, dy
\]

\[
= E[g(X, Y)].
\]

**Example 5.15.** We continue Example 5.14. The conditional expectation of \( X \) given \( Y \) is the random variable

\[
E[X \mid Y] = \frac{1 - Y}{2} =: Z.
\]
The pdf of $Z$ is

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

which is illustrated in Figure 5.2. Note that

$$E[Z] = \int_0^{\frac{1}{2}} 8z^2 \, dz = \frac{1}{3} = E[X],$$

as is expected from the law of iterated expectation. Similarly,

$$E[XY] = E[E[XY \mid Y]]$$

$$= E\left[ \frac{Y(1-Y)}{2} \right]$$

$$= \int_0^1 \frac{y(1-y)}{2} \cdot 2(1-y) \, dy = \frac{1}{12},$$

which agrees with the direct integration computed in Example 5.14.

**Example 5.16.** 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, that is, $N \mid \{P = p\} \sim \text{Binom}(n, p)$. Then, by the law of iterated expectation, we can find

$$E[N] = E[E[N \mid P]]$$

$$= E[nP]$$

$$= nE[P]$$

$$= n \int_0^1 2(1 - p)p \, dp = \frac{1}{3}n,$$

which is much simpler than finding the pmf of $N$ and computing the expectation.

![Figure 5.2. The graph of $f_Z(z)$.](image)
Example 5.17. Let $E[X | Y] = Y^2$ and $Y \sim \text{Unif}[0, 1]$. In this case, we cannot find the pdf of $X$, since we do not know $f_{X|Y}(x|y)$. But using iterated expectation we can still find

$$E[X] = E[E[X | Y]] = E[Y^2] = \int_0^1 y^2 \, dy = \frac{1}{3}.$$  

We define the conditional variance of $X$ given $Y = y$ as the variance of $X$ w.r.t. $f_{X|Y}(x|y)$, i.e.,

$$\text{Var}(X | Y = y) = E[(X - E[X | Y = y])^2 | Y = y]$$

$$= E[X^2 | Y = y] - 2E[X E[X | Y = y] | Y = y] + E[(E[X | Y = y])^2 | Y = y]$$

$$= E[X^2 | Y = y] - (E[X | Y = y])^2.$$  

The random variable $\text{Var}(X | Y)$ is a function of $Y$ that takes on the values $\text{Var}(X | Y = y)$. Its expected value is

$$E[\text{Var}(X | Y)] = E[E[X^2 | Y] - (E[X | Y])^2] = E[X^2] - E[(E[X | Y])^2].$$  

(5.2)  

Since $E(X | Y)$ is a random variable, it has a variance

$$\text{Var}(E(X | Y)) = E[(E[X | Y] - E[E[X | Y]])^2]$$


(5.3)

By adding (5.2) and (5.3), we establish the law of conditional variances:

$$\text{Var}(X) = E[\text{Var}(X | Y)] + \text{Var}(E[X | Y]).$$

5.6 MMSE ESTIMATION

Consider the signal estimation system depicted in Figure 5.3 where the original signal is $X \sim f_X(x)$ and its noisy observation is

$$Y | \{X = x\} \sim f_{Y|X}(y|x).$$

An estimator is a mapping $g : \mathbb{R} \to \mathbb{R}$ that generates an estimate $\hat{X} = g(Y)$ of the original signal. We wish to find an optimal estimator $g^*(y)$ that minimizes the mean square error (MSE)

$$E[(X - \hat{X})^2] = E[(X - g(Y))^2].$$  

(5.4)

![Figure 5.3. The signal estimation system.](image)
The estimator $g^*(y)$ that attains the smallest value of (5.3) is referred to as the minimum mean square error (MMSE) estimator of $X$ given $Y$, and \( \hat{X} = g^*(Y) \) is referred to as the MMSE estimate.

Suppose that there is no observation and let $a^*$ be the MMSE estimate of $X$, that is,

\[
a^* = \arg \min_a E[(X - a)^2].
\]

Then,

\[
a^* = E[X]. \quad (5.5)
\]

In other words, the mean is the optimal summary of $X$ under the mean square error criterion. To prove (5.5), note that for any estimate $a$ of $X$,

\[
E[(X - a)^2] = E[(X - E[X] + E[X] - a)^2]
\]
\[
= E[(X - E[X])^2] + (E[X] - a)^2 + E[X - E[X]](E[X] - a)
\]
\[
= E[(X - E[X])^2] + (E[X] - a)^2
\]
\[
\geq E[(X - E[X])^2]
\]

with equality iff $a = E[X]$.

A similar observation continues to hold with the observation $Y = y$. Let $g^*(y) = E[X \mid Y = y]$. Then, by (5.3), for any estimator $g(y)$,

\[
E[(X - g^*(y))^2 \mid Y = y] \leq E[(X - g(y))^2 \mid Y = y].
\]

Consequently,

\[
E[(X - g^*(Y))^2] \leq E[(X - g(Y))^2]
\]

and \( \hat{X} = g^*(Y) = E[X \mid Y] \) is the MMSE estimate of $X$ given $Y$ with the corresponding MSE

\[
E[\text{Var}(X \mid Y)] = E[(X - E[X \mid Y])^2].
\]

The MMSE estimate \( \hat{X} = E[X \mid Y] \) satisfies the following properties.

1. It is unbiased, i.e.,

\[
E[\hat{X}] = E[X].
\]

2. The estimation error $X - \hat{X}$ is unbiased for every $Y = y$, i.e.,

\[
E[X - \hat{X} \mid Y = y] = 0.
\]

3. The estimation error and the estimate are orthogonal, i.e.,

\[
E[(X - \hat{X})\hat{X}] = E[E[(X - \hat{X})\hat{X} \mid Y]]
\]
\[
= E[E[\hat{X} E[X - \hat{X} \mid Y]]] = 0.
\]

In fact, the estimation error is orthogonal to any function $g(Y)$, i.e.,

\[
E[(X - \hat{X})g(Y)] = 0.
\]
4. By the law of conditional variance \(\text{Var}(X) = \text{Var}({\hat{X}}) + \mathbb{E}[\text{Var}(X | Y)]\), the sum of the variance of the estimate and its MSE is equal to the variance of the signal.

5. If \(X\) and \(Y\) are independent, then \(\hat{X} = \mathbb{E}[X]\), that is, the observation is ignored.

**Example 5.18.** Again let

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

We find the MMSE estimate of \(X\) given \(Y\) and its MSE. We already know that the MMSE estimate is

\[
\mathbb{E}[X | Y] = \frac{1 - Y}{2}
\]

and that the conditional variance is

\[
\text{Var}[X | Y] = \frac{(1 - Y)^2}{12}.
\]

Hence, the MMSE is \(\mathbb{E}[\text{Var}(X | Y)] = 1/24\), compared to \(\text{Var}(X) = 1/18\). The difference is \(\text{Var}(\mathbb{E}[X | Y]) = 1/72\), which is the variance of the estimate.

**Figure 5.4.** The conditional mean and variance of \(X\) given \(Y = y\).

**Example 5.19 (Additive Gaussian noise channel).** Consider a communication channel with input \(X \sim N(\mu, P)\), noise \(Z \sim N(0, N)\), and output \(Y = X + Z\). We assume that \(X\) and \(Z\) are independent. We find the MMSE estimate of \(X\) given \(Y\) and its MSE, i.e., \(\mathbb{E}[X | Y]\) and \(\mathbb{E}[\text{Var}(X | Y)]\). Recall that \(Y | \{X = x\} \sim N(x, N)\) and \(Y \sim N(\mu, P + N)\), that is,

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

and

\[
f_Y(y) = \frac{1}{\sqrt{2\pi(P + N)}} e^{-\frac{(y-\mu)^2}{2(P+N)}}.
\]
5.6 MMSE Estimation

Hence,

\[
f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)} = \frac{1}{\sqrt{2\pi p}} e^{-\frac{(x-y)^2}{2p}} \left( \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-\mu)^2}{2N}} \right) = \frac{1}{\sqrt{2\pi PN}} \exp \left( -\frac{(x - \left( \frac{P}{P+N} y + \frac{N}{P+N} \mu \right))^2}{2\left( \frac{PN}{P+N} \right)} \right),
\]

or equivalently,

\[
X \mid \{Y = y\} \sim N\left( \frac{P}{P+N} y + \frac{N}{P+N} \mu, \frac{PN}{P+N} \right).
\]

Thus,

\[
E[X \mid Y] = \frac{P}{P+N} Y + \frac{N}{P+N} \mu,
\]

which is a convex combination of the observation \(Y\) and the mean \(\mu\) (MMSE estimate without observation), and tends to \(Y\) as \(N \to 0\) and to \(\mu\) as \(N \to \infty\). The corresponding MSE is

\[
E[\text{Var}(X \mid Y)] = E \left[ \frac{PN}{P+N} \right] = \frac{PN}{P+N},
\]

which is less than \(P\), the MSE without the observation \(Y\). Note that the conditional variance \(\text{Var}(X \mid Y)\) is independent of \(Y\).

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.

**Example 5.20.** Let

\[
f(x\mid y) = \begin{cases} ye^{-y^x} & x \geq 0, \ y > 0, \\ 0 & \text{otherwise.} \end{cases}
\]

Then,

\[
E[X \mid Y] = \frac{1}{Y}.
\]

**Remark 5.3.** There can be alternative criteria for measuring goodness of estimators. For example, instead of the MSE criteria in (5.4) that was introduced in the 19th century by Legendre and Gauss, one may measure the mean absolute error (MAE)

\[
E[|X - g(Y)|],
\]

which dates back Boscovich and Laplace in the preceding century. It can be shown that the minimum MAE estimate is the conditional median, that is,

\[
P[X \leq g^*(y) \mid Y = y] \geq 1/2,
\]

\[
P[X \geq g^*(y) \mid Y = y] \geq 1/2.
\]
5.7 LINEAR MMSE ESTIMATION

To find the MMSE estimate, one needs to know the statistics of the signal and the channel, namely, $f_{X,Y}(x, y)$, or at least, $f_{X|Y}(x|y)$, 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., the 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 linear MMSE (LMMSE) 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

$$
E[(X - \hat{X})^2] = E[(X - aY - b)^2].
$$

We show that the LMMSE estimate of $X$ given $Y$ is

$$
\hat{X} = a^*Y + b^* = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}(Y - E[Y]) + E[X]
$$

and its MSE is

$$
E[(X - a^*Y - b^*)^2] = \text{Var}(X) - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(Y)} = (1 - \rho_{X,Y}^2)\sigma_X^2.
$$

First note that for any $a$,

$$
E[(X - aY - b)^2] = E[((X - aY) - b)^2]
$$

is minimized by $b^*(a) = E[X - aY] = E[X] - aE[Y]$. Hence, under this choice, the MSE can be written as a quadratic function in $a$ as

$$
E[(X - aY - b^*(a))^2] = E[((X - E[X]) - a(Y - E[Y]))^2]
$$

$$
= \text{Var}(X) - 2a \text{Cov}(X, Y) + a^2 \text{Var}(Y),
$$

which is minimized at

$$
a^* = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}
$$

with the minimum

$$
\text{Var}(X) - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(Y)}.
$$
Alternatively, we can minimize
\[ J(a, b) = E[(X - aY - b)^2] \]
by finding \((a^*, b^*)\) that satisfies
\[ \frac{\partial}{\partial a} J(a, b) = \frac{\partial}{\partial b} J(a, b) = 0 \]
and by showing that the solution attains the global minimum as illustrated in Figure 5.5.

The linear MMSE estimate \(\hat{X}\) satisfies the following properties:
1. It is unbiased, i.e., \(E(\hat{X}) = E(X)\), which was also true for the nonlinear MMSE estimate.
2. The estimation error and the estimate are orthogonal, i.e.,
   \[ E[(X - \hat{X})\hat{X}] = 0. \]
   In fact, the estimation error is orthogonal to any affine function \(aY + b\), i.e.,
   \[ E[(X - \hat{X})(aY + b)] = 0. \]
3. If \(\rho_{X,Y} = 0\), i.e., \(X\) and \(Y\) are uncorrelated, then the observation is ignored and
   \(\hat{X} = E[X]\).
4. If \(\rho_{X,Y} = \pm 1\), i.e., \((X - E(X))\) and \((Y - E(Y))\) are linearly dependent, then the linear estimate is perfect and \(\hat{X} = X\).

The LMMSE estimate is not, in general, as good as the MMSE estimate.
Example 5.21. Let $Y \sim \text{Unif}[-1, 1]$ and $X = Y^2$. The MMSE estimate of $X$ given $Y$ is $Y^2$, which is perfect. To find the LMMSE estimate we compute

$$E[Y] = 0,$$
$$E[X] = \int_{-1}^{1} \frac{1}{2} y^2 \, dy = \frac{1}{3},$$

and


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

5.8 GEOMETRIC FORMULATION OF ESTIMATION

For both nonlinear and linear MMSE estimation problems we discussed in the previous two sections, we found that the estimation error is orthogonal to the optimal estimate. This orthogonality property is a fundamental characteristic of an optimal estimator that minimizes the MSE among a class of estimators and can be used to find the optimal estimator in a simple geometric argument.

First, we introduce some mathematical background. A vector space $\mathcal{V}$ consists of a set of vectors that are closed under two operations:

- Vector addition: if $v, w \in \mathcal{V}$ then $v + w \in \mathcal{V}$.
- Scalar multiplication: if $a \in \mathbb{R}$ and $v \in \mathcal{V}$, then $av \in \mathcal{V}$.

An inner product is a real-valued operation $v \cdot w$ satisfying these three conditions:

- Commutativity: $v \cdot w = w \cdot v$.
- Linearity: $(av + v) \cdot w = a(v \cdot w) + v \cdot w$.
- Nonnegativity: $v \cdot w \geq 0$ and $v \cdot v = 0$ iff $v = 0$.

A vector space with an inner product is referred to as an inner product space. For example, the Euclidean space

$$\mathbb{R}^n = \{x = (x_1, x_2, \ldots, x_n): x_1, x_2, \ldots, x_n \in \mathbb{R}\}$$

with vector addition

$$x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n),$$

scalar multiplication

$$ax = (ax_1, ax_2, \ldots, ax_n),$$
and dot product
\[ \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i \]
is an inner product space.

The inner product induces generalized notions of length, distance, and angle. The norm of \( \mathbf{v} \in \mathcal{V} \) is defined as \( \|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} \), and can be viewed as the “length” of \( \mathbf{v} \). A metric is a function \( d(\mathbf{v}, \mathbf{w}) \) satisfying the following three conditions:

- **Commutativity:** \( d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v}) \).
- **Nonnegativity:** \( d(\mathbf{v}, \mathbf{w}) \geq 0 \) with equality iff \( \mathbf{v} = \mathbf{w} \).
- **Triangle inequality:** \( d(\mathbf{u}, \mathbf{w}) \leq d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w}) \).

It is easy to verify that \( \|\mathbf{v} - \mathbf{w}\| \) is a metric, and thus can be interpreted as the “distance” between \( \mathbf{v} \) and \( \mathbf{w} \). We say that \( \mathbf{v} \) and \( \mathbf{w} \) are orthogonal (written \( \mathbf{v} \perp \mathbf{w} \)) if \( \mathbf{v} \cdot \mathbf{w} = 0 \). More generally, the “angle” \( \theta \) between \( \mathbf{v} \) and \( \mathbf{w} \) satisfies \( \mathbf{v} \cdot \mathbf{w} = \|\mathbf{v}\| \|\mathbf{w}\| \cos \theta \). For example, the norm (length) of a vector \( \mathbf{x} \) in the Euclidean space is
\[ \|\mathbf{x}\| = \left( \sum_{i=1}^{n} x_i^2 \right)^{1/2}, \]
the distance between the two points represented by \( \mathbf{x} \) and \( \mathbf{y} \) is
\[ \|\mathbf{x} - \mathbf{y}\| = \left( \sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}, \]
and the angle between the two vectors \( \mathbf{x} \) and \( \mathbf{y} \) is
\[ \theta = \arccos \frac{\sum_{i=1}^{n} x_i y_i}{\left( \sum_{i=1}^{n} x_i^2 \right)^{1/2} \left( \sum_{i=1}^{n} y_i^2 \right)^{1/2}}. \]

The Pythagorean theorem holds in an arbitrary inner product space, namely, if \( \mathbf{v} \perp \mathbf{w} \), then
\[ \|\mathbf{v} + \mathbf{w}\|^2 = (\mathbf{v} + \mathbf{w}) \cdot (\mathbf{v} + \mathbf{w}) = (\mathbf{v} \cdot \mathbf{v}) + (\mathbf{w} \cdot \mathbf{w}) + 2(\mathbf{v} \cdot \mathbf{w}) = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2. \]

We say that \( \mathcal{W} \) is a subspace of a vector space \( \mathcal{V} \) if \( \mathcal{W} \subseteq \mathcal{V} \) is itself a vector space (i.e., closed under vector addition and scalar multiplication). A subspace of an inner product space inherits the same inner product and is also an inner product space. We now establish the following simple observation on the distance between a vector in a vector space \( \mathcal{V} \) and its subspace \( \mathcal{W} \).

**Orthogonality principle.** Let \( \mathcal{V} \) be an inner product space and \( \mathcal{W} \) be its subspace. Let \( \mathbf{v} \in \mathcal{V} \). Suppose that there exists \( \mathbf{w}^* \in \mathcal{W} \) such that \( \mathbf{v} - \mathbf{w}^* \) is orthogonal to every \( \mathbf{w} \in \mathcal{W} \). Then,
\[ \mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathcal{W}} \|\mathbf{v} - \mathbf{w}\|. \]
As depicted in Figure 5.6, the orthogonal projection of \( v \) onto \( \mathcal{W} \) (if it exists) is the closest vector of \( v \) in \( \mathcal{W} \). The proof is immediate from the Pythagorean theorem. For any \( w \in \mathcal{V} \), \( (v - w^*) \perp (w^* - w) \) by the orthogonality condition. Hence, \( \|v - w^*\|^2 + \|w^* - w\|^2 = \|v - w\|^2 \) and thus \( \|v - w^*\| \leq \|v - w\| \).

![Figure 5.6. Among all vectors in \( \mathcal{W} \), the orthogonal projection \( w^* \) of \( v \) is the closest.](image)

We now consider the inner product space \( \mathcal{V} \) that consists of all random variables (with finite second moment) on the same probability space, where

- the vector addition (sum) \( V + W \) of random variables \( V \) and \( W \) is a random variable,
- the scalar (constant) multiplication \( aV \) is a random variable, and
- the inner product \( V \cdot W = 0 \) of \( V \) and \( W \) is their correlation \( \mathbb{E}[VW] \) (which satisfies the three inner product axioms).

Fortuitously, two random variables \( V \) and \( W \) are orthogonal, i.e., \( \mathbb{E}[VW] = 0 \), as defined in Section 5.4, if \( V \) and \( W \) are orthogonal as two vectors, i.e., \( V \cdot W = 0 \). Note that the norm of \( V \) is \( \|V\| = \sqrt{\mathbb{E}[V^2]} \).

The goal of MMSE estimation can be now rephrased as follows: Given the vector space \( \mathcal{V} \) of all random variables (or all random variables that are functions of \( X \) and \( Y \)) and a subspace \( \mathcal{W} \) of estimators, find \( \hat{X} \) that is closest to \( X \), that is, the mean square error \( \|\hat{X} - X\|^2 \) is the smallest.

**Example 5.22 (MMSE estimator).** Let \( \mathcal{W} \) be the space of all functions \( g(Y) \) with finite second moment. It can be easily verified that it is an inner product space. We already know that the MMSE estimate \( \hat{X} = g^*(Y) = \mathbb{E}[X|Y] \) we found in Section 5.6 has the property that the error \( \hat{X} - X \) is orthogonal to every \( g(Y) \). Hence, it minimizes the MSE among all functions of \( Y \).

**Example 5.23 (Mean).** Let \( \mathcal{W} \) be the set of all constants \( a \in \mathbb{R} \). Once again it is a valid subspace. Since \( X - \mathbb{E}[X] \) is orthogonal to \( \mathcal{W} \), i.e., \( \mathbb{E}[(X - \mathbb{E}[X])a] = 0 \) for every \( a \), \( \hat{X} = a^* = \mathbb{E}[X] \) minimizes the MSE among all constants.

**Example 5.24 (LMMSE estimator).** Let \( \mathcal{W} \) be the subspace that consists all functions of
the form $aY + b$. Since $X - (a^* Y + b^*)$, where $a^*$ and $b^*$ are given in (5.6), is orthogonal to any $aY + b$, $\hat{X} = a^* Y + b^*$ minimizes the MSE among all affine functions of $Y$.

We shall later apply this orthogonality principle to find MMSE estimators in more general subspaces such as linear combinations of multiple random variables and linear filters of random processes.

### 5.9 Jointly Gaussian Random Variables

We say that two random variables are **jointly Gaussian** if their joint pdf is of the form

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

Note that this pdf is a function only of $\mu_X$, $\mu_Y$, $\sigma_X^2$, $\sigma_Y^2$, and $\rho_{X,Y}$. Consistent with our notation, these parameters are indeed $E[X]$, $E[Y]$, $\text{Var}(X)$, $\text{Var}(Y)$, and the correlation coefficient of $X$ and $Y$. In Lecture #6, we shall define jointly Gaussian random variables in a more general way.

**Example 5.25.** Consider the additive Gaussian noise channel in Example 5.19 where $X \sim N(\mu, P)$ and $Z \sim N(0, N)$ are independent and $Y = X + Z$. Then the pair $X$ and $Z$, the pair $X$ and $Y$, and the pair $Y$ and $Z$ are jointly Gaussian.

If $X$ and $Y$ are jointly Gaussian, contours of equal joint pdf are ellipses defined by the quadratic equation

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

The orientation of the major axis of these ellipses is

$$\theta = \frac{1}{2} \arctan \left( \frac{2\rho_{X,Y} \sigma_X \sigma_Y}{\sigma_X^2 - \sigma_Y^2} \right).$$

Figure 5.7 shows a few examples of the joint pdf.

Jointly Gaussian random variables $X$ and $Y$ satisfy the following properties.

1. They are marginally Gaussian, i.e.,
   $$X \sim N(\mu_X, \sigma_X^2) \quad \text{and} \quad Y \sim N(\mu_Y, \sigma_Y^2).$$

2. The conditional pdf is Gaussian, i.e.,
   $$X \mid \{Y = y\} \sim N\left( \frac{\rho_{X,Y} \sigma_X}{\sigma_Y} (y - \mu_Y) + \mu_X, (1 - \rho_{X,Y}^2) \sigma_X^2 \right),$$
   which shows that the MMSE estimate is linear.
\[ \sigma_X = 1, \sigma_Y = 1, \rho_{X,Y} = 0 \]

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

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

Figure 5.7. Joint pdfs of jointly Gaussian random variables.
3. If $X$ and $Y$ are jointly Gaussian and uncorrelated, i.e., $\rho_{X,Y} = 0$, then they are also independent.

The converse to the first property is not necessarily true, that is, Gaussian marginals do not necessarily mean that the random variables are jointly Gaussian.

**Example 5.26.** Let $X \sim \mathcal{N}(0, 1)$ and

$$Z = \begin{cases} +1 & \text{w.p. } 1/2, \\ -1 & \text{w.p. } 1/2 \end{cases}$$

be independent and let $Y = XZ$. Clearly, $Y \sim \mathcal{N}(0, 1)$. However, $X$ and $Y$ do not have a joint pdf. Using delta functions, “$f_{X,Y}(x, y)$” has the form shown in Figure 5.8. Note that $X$ and $Y$ are uncorrelated, but not independent. This does not contradict the third property since $X$ and $Y$ are not jointly Gaussian.

**Figure 5.8.** The “joint pdf” of $X$ and $Y$.

**PROBLEMS**

5.1. *Inequalities.* Label each of the following statements with $=, \leq,$ or $\geq$. Justify each answer.

(a) $1/E[X^2]$ vs. $E(1/X^2)$.

(b) $(E[X])^2$ vs. $E[X^2]$.

(c) $\text{Var}(X)$ vs. $\text{Var}(E[X|Y])$.

(d) $E[X^2]$ vs. $E[(E[X|Y])^2]$.

5.2. *Cauchy–Schwartz inequality.*
(a) Prove the following inequality: \((E[XY])^2 \leq E[X^2] E[Y^2]\). (Hint: Use the fact that for any real \(t\), \(E[(X + tY)^2] \geq 0\).)

(b) Prove that equality holds if and only if \(X = cY\) for some constant \(c\). Find \(c\) in terms of the second moments of \(X\) and \(Y\).

(c) Use the Cauchy–Schwartz inequality to show the correlation coefficient satisfies \(|\rho_{X,Y}| \leq 1\).

(d) Prove the triangle inequality: \(\sqrt{E[(X + Y)^2]} \leq \sqrt{E[X^2]} + \sqrt{E[Y^2]}\).

5.3. Two envelopes. An amount \(A\) is placed in one envelope and the amount \(2A\) is placed in another envelope. The amount \(A\) is fixed but unknown to you. The envelopes are shuffled and you are given one of the envelopes at random. Let \(X\) denote the amount you observe in this envelope. Designate by \(Y\) the amount in the other envelope. Thus

\[
(X, Y) = \begin{cases} 
(A, 2A), & \text{w.p. } 1/2, \\
(2A, A), & \text{w.p. } 1/2.
\end{cases}
\]

You may keep the envelope you are given, or you can switch envelopes and receive the amount in the other envelope.

(a) Find \(E[X]\) and \(E[Y]\).

(b) Find \(E[X/Y]\) and \(E[Y/X]\).

(c) Suppose you switch. What is the expected amount you receive?

5.4. Mean and variance. Let \(X\) and \(Y\) be random variables with joint pdf

\[
f_{X,Y}(x, y) = \begin{cases} 
1 & \text{if } |x| + |y| \leq 1/\sqrt{2}, \\
0 & \text{otherwise}.
\end{cases}
\]

Define the random variable \(Z = |X| + |Y|\). Find the mean and variance of \(Z\) without first finding the pdf of \(Z\).

5.5. Tall trees. Suppose that the average height of trees on campus is 20 feet. Argue that no more than half of the tree population is taller than 40 feet.

5.6. Let \(X\) and \(Y\) have correlation coefficient \(\rho_{X,Y}\).

(a) What is the correlation coefficient between \(X\) and \(3Y\)?

(b) What is the correlation coefficient between \(2X\) and \(-5Y\)?

5.7. Random phase signal. Let \(Y(t) = \sin(\omega t + \Theta)\) be a sinusoidal signal with random phase \(\Theta \sim \text{Unif}[-\pi, \pi]\). Assume here that \(\omega\) and \(t\) are constants. Find the mean and variance of \(Y(t)\). Do they depend on \(t\)?

5.8. Coin tosses. A coin with bias \(p\) is tossed independently until two heads or two tails come up in a row. Find the expected value of the number of tosses \(X\).
5.9. *Iterated expectation.* Let \( \Lambda \) and \( X \) be two random variables with

\[
\Lambda \sim f_{\Lambda}(\lambda) = \begin{cases} 
\frac{5}{3} \lambda^2, & 0 \leq \lambda \leq 1 \\
0, & \text{otherwise,}
\end{cases}
\]

and \( X | \{ \Lambda = \lambda \} \sim \text{Exp}(\lambda) \). Find \( E(X) \).

5.10. *Sum of packet arrivals.* Consider a network router with two types of incoming packets, wireline and wireless. Let the random variable \( N_1(t) \) denote the number of wireline packets arriving during time \( (0, t] \) and let the random variable \( N_2(t) \) denote the number of wireless packets arriving during time \( (0, t] \). Suppose \( N_1(t) \) and \( N_2(t) \) are independent Poisson with pmfs

\[
P\{N_1(t) = n\} = \frac{\lambda_1 t^n}{n!} e^{-\lambda_1 t} \quad \text{for } n = 0, 1, 2, \ldots
\]

\[
P\{N_2(t) = k\} = \frac{\lambda_2 t^k}{k!} e^{-\lambda_2 t} \quad \text{for } k = 0, 1, 2, \ldots
\]

Let \( N(t) = N_1(t) + N_2(t) \) be the total number of packets arriving at the router during time \( (0, t] \).

(a) Find the mean \( E(N(t)) \) and variance \( \text{Var}(N(t)) \) of the total number of packet arrivals.

(b) Find the pmf of \( N(t) \).

(c) Let the random variable \( Y \) be the time to receive the first packet of either type. Find the pdf of \( Y \).

(d) What is the probability that the first received packet is wireless?

5.11. *Conditioning on an event.* Let \( X \) be a r.v. with pdf

\[
f_X(x) = \begin{cases} 
2(1 - x) & \text{for } 0 \leq x \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

and let the event \( A = \{ X \geq 1/3 \} \). Find \( f_{X|A}(x) \), \( E(X|A) \), and \( \text{Var}(X|A) \).

5.12. *Jointly Gaussian random variables.* Let \( X \) and \( Y \) be jointly Gaussian random variables with pdf

\[
f_{X,Y}(x, y) = \frac{1}{\pi \sqrt{3/4}} e^{-\frac{1}{2} \left( 4x^2/3 + 16y^2/3 + 32x^2y^2/3 - 8x - 16y + 16 \right)}.
\]

Find \( E(X), E(Y), \text{Var}(X), \text{Var}(Y) \), and \( \text{Cov}(X, Y) \).

5.13. *Neural net.* Let \( Y = X + Z \), where the signal \( X \sim U[-1, 1] \) and noise \( Z \sim \mathcal{N}(0, 1) \) are independent.
(a) Find the function \( g(y) \) that minimizes

\[
\text{MSE} = \mathbb{E} \left[ (\text{sgn}(X) - g(Y))^2 \right],
\]

where

\[
\text{sgn}(x) = \begin{cases} 
-1 & x \leq 0 \\
+1 & x > 0.
\end{cases}
\]

(b) Plot \( g(y) \) vs. \( y \).

5.14. **Additive shot noise channel.** Consider an additive noise channel \( Y = X + Z \), where the signal \( X \sim \mathcal{N}(0, 1) \), and the noise \( Z|X = x \sim \mathcal{N}(0, x^2) \), i.e., the noise power of increases linearly with the signal squared.

(a) Find \( E(Z^2) \).

(b) Find the best linear MSE estimate of \( X \) given \( Y \).

5.15. **Additive uniform noise channel.** Let the signal

\[
X = \begin{cases} 
+1, & \text{with probability } \frac{1}{2}, \\
-1, & \text{with probability } \frac{1}{2},
\end{cases}
\]

and the noise \( Z \sim \text{Unif}[-2, 2] \) be independent random variables. Their sum \( Y = X + Z \) is observed. Find the minimum MSE estimate of \( X \) given \( Y \) and its MSE.

5.16. **Estimation vs. detection.** Let the signal

\[
X = \begin{cases} 
+1, & \text{with probability } \frac{1}{2}, \\
-1, & \text{with probability } \frac{1}{2},
\end{cases}
\]

and the noise \( Z \sim \text{Unif}[-2, 2] \) be independent random variables. Their sum \( Y = X + Z \) is observed.

(a) Find the best MSE estimate of \( X \) given \( Y \) and its MSE.

(b) Now suppose we use a decoder to decide whether \( X = +1 \) or \( X = -1 \) so that the probability of error is minimizied. Find the optimal decoder and its probability of error. Compare the optimal decoder’s MSE to the minimum MSE.

5.17. **Linear estimator.** Consider a channel with the observation \( Y = XZ \), where the signal \( X \) and the noise \( Z \) are uncorrelated Gaussian random variables. Let \( E[X] = 1, E[Z] = 2, \sigma_X^2 = 5, \) and \( \sigma_Z^2 = 8 \).

(a) Find the best MSE linear estimate of \( X \) given \( Y \).

(b) Suppose your friend from Caltech tells you that he was able to derive an estimator with a lower MSE. Your friend from UCLA disagrees, saying that this is not possible because the signal and the noise are Gaussian, and hence the best linear MSE estimator will also be the best MSE estimator. Could your UCLA friend be wrong?
5.18. Additive-noise channel with path gain. Consider the additive noise channel shown in the figure below, where $X$ and $Z$ are zero mean and uncorrelated, and $a$ and $b$ are constants.

$$Y = b(aX + Z)$$

Find the MMSE linear estimate of $X$ given $Y$ and its MSE in terms only of $\sigma_X$, $\sigma_Z$, $a$, and $b$.

5.19. Worst noise distribution. Consider an additive noise channel $Y = X + Z$, where the signal $X \sim \mathcal{N}(0, P)$ and the noise $Z$ has zero mean and variance $N$. Assume $X$ and $Z$ are independent. Find a distribution of $Z$ that maximizes the minimum MSE of estimating $X$ given $Y$, i.e., the distribution of the worst noise $Z$ that has the given mean and variance. You need to justify your answer.

5.20. Image processing. A pixel signal $X \sim \mathcal{U}[-k, k]$ is digitized to obtain

$$\tilde{X} = i + \frac{1}{2}, \text{ if } i < X \leq i + 1, \ i = -k, -k + 1, \ldots, k - 2, k - 1.$$ 

To improve the visual appearance, the digitized value $\tilde{X}$ is dithered by adding an independent noise $Z$ with mean $E(Z) = 0$ and variance $\text{Var}(Z) = N$ to obtain $Y = \tilde{X} + Z$.

(a) Find the correlation of $X$ and $Y$.

(b) Find the best linear MSE estimate of $X$ given $Y$. Your answer should be in terms only of $k$, $N$, and $Y$.

5.21. Orthogonality. Let $\hat{X}$ be the minimum MSE estimate of $X$ given $Y$.

(a) Show that for any function $g(y)$, $E((X - \hat{X})g(Y)) = 0$, i.e., the error $(X - \hat{X})$ and $g(Y)$ are orthogonal.

(b) Show that

$$\text{Var}(X) = E(\text{Var}(X|Y)) + \text{Var}(\hat{X}).$$

Provide a geometric interpretation for this result.

5.22. Difference from sum. Let $X$ and $Y$ be two random variables. Let $Z = X + Y$ and let $W = X - Y$. Find the best linear estimate of $W$ given $Z$ as a function of $E(X)$, $E(Y)$, $\sigma_X$, $\sigma_Y$, $\rho_{XY}$ and $Z$. 
5.23. **Nonlinear and linear estimation.** Let $X$ and $Y$ be two random variables with joint pdf

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

(a) Find the MMSE estimator of $X$ given $Y$.

(b) Find the corresponding MSE.

(c) Find the pdf of $Z = E(X|Y)$.

(d) Find the linear MMSE estimator of $X$ given $Y$.

(e) Find the corresponding MSE.

5.24. **Additive-noise channel with signal dependent noise.** Consider the channel with correlated signal $X$ and noise $Z$ and observation $Y = 2X + Z$, where

$$\mu_X = 1, \quad \mu_Z = 0, \quad \sigma_X^2 = 4, \quad \sigma_Z^2 = 9, \quad \rho_{X,Z} = -\frac{3}{8}.$$

Find the best MSE linear estimate of $X$ given $Y$. 