UC Berkeley Compendium

Frank Lin

March 30, 2019

Contents

Part I

Mathematics

Chapter 1

Linear Algebra

1.1 Linear Equations

1.1.1 Systems of Linear Equations

Definition 1. A **linear equation** in the variables $x_1, \ldots, x_n$ is an equation that can be written in the form $a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$ where b and the coefficients $a_1, \ldots, a_n$ are real or complex numbers and the subscript n can be any positive integer.

Definition 2. A **linear system** is a collection of one or more linear equations involving the same variables.

Definition 3. A **solution** of the system is a list of numbers $(s_1, s_2, \ldots, s_n)$ that makes each equation a true statement when the values $s_1, \ldots, s_n$ are substituted for $x_1, \ldots, x_n$, respectively. The set of all possible solutions is called the **solution set** of the linear system.

Definition 4. Two linear systems are called **equivalent** if they have the same solution set. That is, each solution of the first system is a solution of the second system, and vice versa.

A system of linear equations has either

1. no solution, or
2. exactly one solution, or
3. infinitely many solutions

Definition 5. A system of linear equations is said to be **consistent** if it has either one solution or infinitely many solutions. A system is **inconsistent** if it has no solution.

The essential information of a linear system can be recorded compactly in a rectangular array called a matrix. For example, given the following linear system

$$\begin{aligned} x_1 - 5x_2 + 4x_3 &= 0 \\ 2x_1 - 7x_2 + 3x_2 &= -2 \\ -2x_1 + x_2 + 7x_2 &= -1 \end{aligned}$$

we have the following **coefficient matrix** of the system

$$\begin{bmatrix} 1 & -5 & 4 \\ 2 & -7 & 3 \\ -2 & 1 & 7 \end{bmatrix}$$

and the following **augmented matrix** of the system

$$\left[\begin{array}{ccc|c} 1 & -5 & 4 & 0 \\ 2 & -7 & 3 & -2 \\ -2 & 1 & 7 & -1 \end{array}\right]$$

The **size** of a matrix tells how many rows and columns it has. For example, the augmented matrix above has 3 rows and 4 columns and is called a 3×4 matrix. If m and n are positive numbers, an $m \times n$ matrix is a rectangular array of numbers with m rows and n columns.

The following is an algorithm for solving linear systems. The basic strategy is to replace one system with an equivalent system, i.e. one with the same solution set, that is easier to solve.

The following **elementary row operations** are used to simplify a linear system:

1. Replace one row by the sum of itself and a multiple of another row.
2. Interchange two rows.
3. Multiply all entries in a row by a nonzero constant.

It is important to note that row operations are reversible.

Definition 6. Two matrices are called **row equivalent** if there is a sequence of elementary row operations that transforms one matrix into the other.

If the augmented matrices of two linear systems are row equivalent, then the two systems have the same solution set.

For each particular linear system, we ask two questions:

1. Does at least one solution exist?
2. If a solution exists, is it unique?

1.1.2 Row Reduction and Echelon Forms

Definition 7. A rectangular matrix is in **row echelon form** if it has the following three properties:

1. All nonzero rows are above any rows of all zeros.
2. Each leading entry of a row is in a column to the right of the leading entry of the row above it.
3. All entries in a column below a leading entry, the leftmost nonzero entry in a nonzero row, are zeros.

If a matrix in row echelon form additionally satisfies the following conditions, then it is in **reduced (row) echelon form**:

4. The leading entry in each nonzero row is 1.
5. Each leading 1 is the only nonzero entry in its column.

An echelon matrix (resp. reduced echelon matrix) is one that is in echelon form (resp. reduced echelon form).

Theorem 1. Each matrix is row equivalent to one and only one reduced echelon matrix.

Definition 8. A **pivot position** in a matrix A is a location in A that corresponds to a leading 1 in the reduced echelon form of A. A **pivot column** is a column of A that contains a pivot position. A **pivot** is a nonzero number in a pivot position that is used to create zeros via row operations.

The following is the **row reduction algorithm**, consisting of four steps that produces a matrix in echelon form. A fifth step produces a matrix in reduced echelon form.

1. Begin with the leftmost nonzero column. This is a pivot column and the pivot position is at the top.
2. Select a nonzero entry in the pivot column as a pivot. If necessary, interchange rows to move this entry into the pivot position.
3. Use row replacement operations to create zeros in all positions below the pivot.
4. Ignoring the row containing the pivot position and all rows, if any, above it, repeatedly apply the previous three steps to the remaining submatrix until there are no more nonzero rows to modify.
5. Beginning with the rightmost pivot and working upward and to the left, create zeros above each pivot. If a pivot is not 1, make it so by a scaling operation.

The row reduction algorithm leads directly to an explicit description of the solution set of a linear system when the algorithm is applied to the augmented matrix of the system.

Suppose we apply the row reduction algorithm to the augmented matrix of a linear system and we obtain the following equivalent reduced echelon form:

$$\begin{bmatrix} 1 & 0 & -5 & 1 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Then we have the following associated system of equations:

$$\begin{aligned} x_1 - 5x_3 &= 1 \\ x_2 + x_3 &= 4 \\ 0 &= 0 \end{aligned}$$

The variables x_1 and x_2 corresponding to pivot columns in the matrix are called **basic variables** and the other variable x_3 is called a **free variable**.

The parametric description of the solution set is the following:

$$\begin{aligned} x_1 &= 1 + 5x_3 \\ x_2 &= 4 - x_3 \\ x_3 &\text{ is free} \end{aligned}$$

Whenever a system is consistent, the solution set can be described explicitly by solving the reduced system of equations for the basic variables in terms of the free variables.

Theorem 2. A linear system is consistent if and only if the rightmost column of the augmented matrix is not a pivot column, that is, if and only if an echelon form of the augmented matrix has no row of the form $\begin{bmatrix} 0 & \cdots & 0 & b \end{bmatrix}$ with b nonzero. If a linear system is consistent, then the solution set contains either (i) a unique solution, when there are no free variables, or (ii) infinitely many solutions, when there is at least one free variable.

The following procedure outlines how to find and describe all solutions of a linear system:

1. Write the augmented matrix of the system.
2. Use the row reduction algorithm to obtain an equivalent augmented matrix in echelon form. Decide whether the system is consistent. If there is no solution, stop; otherwise, go to the next step.
3. Continue row reduction to obtain the reduced echelon form.
4. Write the system of equations corresponding to the matrix obtained above.
5. Rewrite each nonzero equation so that its one basic variable is expressed in terms of any free variables appearing in the equation.

1.1.3 Vector Equations

Definition 9. A matrix with only one column is called a **(column) vector**.

The set of all vectors with two entries is denoted $\mathbb{R}^2$. Two vectors in $\mathbb{R}^2$ are **equal** if and only if their corresponding entries are equal. Given two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^2$, their **sum** is the vector $\mathbf{u}+\mathbf{v}$ obtained by adding corresponding entries of $\mathbf{u}$ and $\mathbf{v}$. The **scalar multiple** of $\mathbf{u}$ by c is the vector $c\mathbf{u}$. The following are examples:

$$\begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 1+2 \\ -2+5 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

$$5 \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 15 \\ -5 \end{bmatrix}$$

Definition 10. The number c in $c\mathbf{u}$ is called a **scalar**.

Considering a rectangular coordinate system in the plane, we can identify geometric points (a, b) with the column vector $\begin{bmatrix} a \\ b \end{bmatrix}$. Similarly, vectors in $\mathbb{R}^3$, 3×1 column matrices with three entries, are represented geometrically by points in a three-dimensional coordinate space.

Definition 11. If n is a positive integer, $\mathbb{R}^n$ denotes the collection of all lists of n real numbers. The vector whose entries are all zero is called the **zero vector** and is denoted $\mathbf{0}$. Equality of vectors in $\mathbb{R}^n$ and operations of scalar multiplication and vector addition in $\mathbb{R}^n$ are defined entry by entry just as in $\mathbb{R}^2$.

The following are algebraic properties of $\mathbb{R}^n$. For all $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in $\mathbb{R}^n$ and all scalars c and d:

1. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
2. $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$
3. $\mathbf{u} + \mathbf{0} = \mathbf{0} + \mathbf{u} = \mathbf{u}$
4. $\mathbf{u} + (-\mathbf{u}) = -\mathbf{u} + \mathbf{u} = \mathbf{0}$
5. $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$
6. $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$
7. $c(d\mathbf{u}) = (cd)\mathbf{u}$
8. $1\mathbf{u} = \mathbf{u}$

Definition 12. Given vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ in $\mathbb{R}^n$ and given scalars $c_1, c_2, \ldots, c_p$, the vector $\mathbf{y} = c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p$ is called a **linear combination** of $\mathbf{v}_1, \ldots, \mathbf{v}_p$ with **weights** $c_1, \ldots, c_p$.

A vector equation $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$ has the same solution set as the linear system whose augmented matrix is $\begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n & \mathbf{b}\end{bmatrix}$. In particular, $\mathbf{b}$ can be generated by a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_n$ if and only if there exists a solution to the linear system corresponding to the above matrix.

One of the key ideas in linear algebra is to study the set of all vectors that can be generated or written as a linear combination of a fixed set $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ of vectors.

Definition 13. If $\mathbf{v}_1, \ldots, \mathbf{v}_p$ are in $\mathbb{R}^n$, then the set of all linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_p$ is denoted by $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ and is called the **subset of $\mathbb{R}^n$ spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_p$**. That is, $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is the collection of all vectors that can be written in the form $c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p$ with $c_1, \ldots, c_p$ scalars.

Asking whether a vector $\mathbf{b}$ is in $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ amounts to asking whether the vector equation $x_1\mathbf{v}_1 + \cdots + x_p\mathbf{v}_p = \mathbf{b}$ has a solution.

1.1.4 The Matrix Equation $A\mathbf{x} = \mathbf{b}$

A fundamental idea in linear algebra is to view a linear combination of vectors as the product of a matrix and a vector.

Definition 14. If A is an $m \times n$ matrix, with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$, and if $\mathbf{x}$ is in $\mathbb{R}^n$, then the **product of A and $\mathbf{x}$**, denoted by $A\mathbf{x}$, is the linear combination of the columns of A using the corresponding entries in $\mathbf{x}$ as weights, i.e.

$$A\mathbf{x} = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{bmatrix}\begin{bmatrix}x_1 \\ \vdots \\ x_n\end{bmatrix} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$$

Note that $A\mathbf{x}$ is defined only if the number of columns of A is equal to the number of entries in $\mathbf{x}$.

Theorem 3. If A is an $m \times n$ matrix, with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$, and if $\mathbf{b}$ is in $\mathbb{R}^n$, the matrix equation $A\mathbf{x} = \mathbf{b}$ has the same solution set as the vector equation $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$ which, in turn, has the same solution as the system of linear equations whose augmented matrix is $\begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n & \mathbf{b}\end{bmatrix}$.

This theorem allows us to view a system of linear equations in three different but equivalent ways: as a matrix equation, as a vector equation, and as a system of linear equations.

The equation $A\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b}$ is a linear combination of the columns of A.

Definition 15. A set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ in $\mathbb{R}^m$ **spans** $\mathbb{R}^m$ if every vector in $\mathbb{R}^m$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_p$, i.e. $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\} = \mathbb{R}^m$.

Theorem 4. Let A be an $m \times n$ matrix. Then the following statements are logically equivalent:

1. For each $\mathbf{b}$ in $\mathbb{R}^m$, the equation $A\mathbf{x} = \mathbf{b}$ has a solution.
2. Each $\mathbf{b}$ in $\mathbb{R}^m$ is a linear combination of the columns of A.
3. The columns of A span $\mathbb{R}^m$.
4. A has a pivot position in every row.

If the product $A\mathbf{x}$ is defined, then the i^{th} entry in $A\mathbf{x}$ is the sum of the products of corresponding entries from row i of A and from the vector $\mathbf{x}$.

Definition 16. The $n \times n$ matrix with 1s on the diagonal and 0s elsewhere is called an **identity matrix** and is denoted by I_n.

Theorem 5. If A is an $m \times n$ matrix, $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^n$ and c is a scalar, then

1. $A(\mathbf{u} + \mathbf{v}) = A\mathbf{u} + A\mathbf{v}$
2. $A(c\mathbf{u}) = c(A\mathbf{u})$

1.1.5 Solution Sets of Linear Systems

Solution sets of linear systems are important objects of study in linear algebra.

Definition 17. A system of linear equations is said to be **homogeneous** if it can be written in the form $A\mathbf{x} = \mathbf{0}$ where A is an $m \times n$ matrix and $\mathbf{0}$ is the zero vector in $\mathbb{R}^m$. Such a solution $A\mathbf{x} = \mathbf{0}$ always has at least one solution, namely $\mathbf{x} = \mathbf{0}$. The zero solution is called the **trivial solution**. A nonzero vector $\mathbf{x}$ that satisfies $A\mathbf{x} = \mathbf{0}$ is called a **nontrivial solution**.

The homogeneous equation $A\mathbf{x} = \mathbf{0}$ has a nontrivial solution if and only if the equation has at least one free variable.

The solution set of a homogeneous equation $A\mathbf{x} = \mathbf{0}$ can always be expressed explicitly as $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ for suitable vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$.

When a nonhomogeneous linear equation has many solutions, the general solution can be written in parametric vector form as one vector plus an arbitrary combination of vectors that satisfy the corresponding homogeneous system.

Theorem 6. Suppose the equation $A\mathbf{x} = \mathbf{b}$ is consistent for some given $\mathbf{b}$, and let $\mathbf{p}$ be a solution. Then the solution set of $A\mathbf{x} = \mathbf{b}$ is the set of all vectors of the form $\mathbf{w} = \mathbf{p} + \mathbf{v}_h$ where $\mathbf{v}_h$ is any solution of the homogeneous equation $A\mathbf{x} = \mathbf{0}$.

We can write a solution set of a consistent system in parametric vector form using the following algorithm:

1. Row reduce the augmented matrix to reduced echelon form.
2. Express each basic variable in terms of any free variables appearing in an equation.
3. Write a typical solution $\mathbf{x}$ as a vector whose entries depend on the free variables, if any.
4. Decompose $\mathbf{x}$ into a linear combination of vectors using the free variables as parameters.

1.1.6 Linear Independence

We can also look at homogeneous equations from a different perspective by writing them as vector equations.

Definition 18. An indexed set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ in $\mathbb{R}^n$ is said to be **linearly independent** if the vector equation

$$x_1\mathbf{v}_1 + \cdots + x_p\mathbf{v}_p = \mathbf{0}$$

has only the trivial solution. The set $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is said to be if there exist weights $c_1, \ldots, c_p$, not all zero, such that

$$c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p = \mathbf{0}$$

This is called a **linear dependence relation** among $\mathbf{v}_1, \ldots, \mathbf{v}_p$ when the weights are not all zero.

Suppose that we begin with a matrix $A = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{bmatrix}$ instead of vectors. The matrix equation $A\mathbf{x} = \mathbf{0}$ can be written as $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0}$. Each linear dependence relation among the columns of A corresponds to a nontrivial solution of $A\mathbf{x} = \mathbf{0}$.

The columns of a matrix A are linearly independent if and only if the equation $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.

A set containing only one vector $\mathbf{v}$ is linearly independent if and only if $\mathbf{v}$ is not the zero vector. A set of two vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$ is linearly dependent if at least one of the vectors is a multiple of the other. The set is linearly independent if and only if neither of the vectors is a multiple of other.

Theorem 7. An indexed set $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ of two or more vectors is linearly dependent if and only if at least one of the vectors in S is a linear combination of the others. In fact, if S is linearly dependent and $\mathbf{v}_1 \neq \mathbf{0}$, then some $\mathbf{v}_j$ (with $j > 1$) is a linear combination of the preceding vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$.

Theorem 8. If a set contains more vectors than there are entries in each vector, then the set is linearly dependent. That is, any set $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ in $\mathbb{R}^n$ is linearly dependent if $p > n$.

Theorem 9. If a set $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ in $\mathbb{R}^n$ contains the zero vector, then the set is linearly dependent.

1.1.7 Linear Transformations

Definition 19. A **transformation** T from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a rule that assigns to each vector $\mathbf{x}$ in $\mathbb{R}^n$ a vector $T(\mathbf{x})$ in $\mathbb{R}^m$. Th set $\mathbb{R}^n$ is called the **domain** of T and $\mathbb{R}^m$ is called the **codomain** of T. For $\mathbf{x}$ in $\mathbb{R}^n$, the vector $T(\mathbf{x})$ in $\mathbb{R}^m$ is called the **image** of $\mathbf{x}$. The set of all images $T(\mathbf{x})$ is called the **range** of T.

In matrix transformations, for each $\mathbf{x}$ in $\mathbb{R}^n$, $T(\mathbf{x})$ is computed as $A\mathbf{x}$, where A is an $m \times n$ matrix. We denote such a matrix transformation by $\mathbf{x} \mapsto A\mathbf{x}$. We note that the domain of T is $\mathbb{R}^n$ when A has n columns and the codomain of T is $\mathbb{R}^m$ when each column of A has m entries. The range of T is the set of all linear combinations of the columns of A.

Definition 20. A transformation T is **linear** if

1. $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ for all $\mathbf{u}$, $\mathbf{v}$ in the domain of T.
2. $T(c\mathbf{u}) = cT(\mathbf{u})$ for all scalars c and all $\mathbf{u}$ in the domain of T.

Linear transformations preserve the operations of vector addition and scalar multiplication.

If T is a linear transformation, then $T(\mathbf{0}) = \mathbf{0}$ and $T(c\mathbf{u} + d\mathbf{v}) = cT(\mathbf{u}) + dT(\mathbf{v})$ for all vectors $\mathbf{u}, \mathbf{v}$ in the domain of T and all scalars c, d.

Repeated application of the above produces a useful generalization:

$$T(c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p) = c_1T(\mathbf{v}_1) + \cdots + c_pT(\mathbf{v}_p)$$

1.1.8 The Matrix of a Linear Transformation

Theorem 10. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then there exists a unique matrix A such that $T(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$. In fact, A is the $m \times n$ matrix whose j^{th} column of the identity matrix in $\mathbb{R}^n$:

$$A = \begin{bmatrix} T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \end{bmatrix}$$

The matrix A is called the standard matrix for the linear transformation T.

Definition 21. A mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ is said to be **onto** $\mathbb{R}^m$ if each $\mathbf{b}$ in $\mathbb{R}^m$ is the image of at least one $\mathbf{x}$ in $\mathbb{R}^n$. A mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ is said to be **one-to-one** if each $\mathbf{b}$ in $\mathbb{R}^m$ is the image of at most one $\mathbf{x}$ in $\mathbb{R}^n$.

Theorem 11. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then T is one-to-one if and only if the equation $T(\mathbf{x}) = \mathbf{0}$ has only the trivial solution.

Theorem 12. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation and let A be the standard matrix for T. Then

1. T maps $\mathbb{R}^n$ onto $\mathbb{R}^m$ if and only if the columns of A span $\mathbb{R}^m$.
2. T is one-to-one if and only if the columns of A are linearly independent.

1.2 Matrix Algebra

1.2.1 Matrix Operations

Definition 1. If A is an $m \times n$ matrix, then the scalar entry in the i^{th} row and j^{th} column of A is denoted by a_{ij}.

Definition 2. The **diagonal entries** in an $m \times n$ matrix A are $a_{11}, a_{22}, \ldots$ and they form the **main diagonal** of A. A **diagonal matrix** is a square $n \times n$ matrix whose nondiagonal entries are zero, e.g. the $n \times n$ identity matrix I_n.

Definition 3. An $m \times n$ matrix whose entries are all zero is a **zero matrix**.

We say two matrices are **equal** if they have the same size and their corresponding columns (and hence, entries) are equal. If A and B are $m \times n$ matrices, then the **sum** $A + B$ is the $m \times n$ matrix whose columns are the sums of the corresponding columns in A and B. The following are examples:

$$\begin{bmatrix} 4 & 0 & 5 \\ -1 & 3 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & 7 \end{bmatrix} = \begin{bmatrix} 5 & 1 & 6 \\ 2 & 8 & 9 \end{bmatrix}$$

$$2B = 2\begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & 7 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 \\ 6 & 10 & 14 \end{bmatrix}$$

Theorem 1. Let A, B, and C be matrices of the same size and let r and s be scalars. Then

1. $A + B = B + A$
2. $(A + B) + C = A + (B + C)$
3. $A + 0 = A$
4. $r(A + B) = rA + rB$
5. $(r + s)A = rA + sA$
6. $r(sA) = (rs)A$

Because of the associative property of addition, we can simply write $A + B + C$ for the sum of three matrices.

Definition 4. If A is an $m \times n$ matrix and B is an $n \times p$ matrix with columns $\mathbf{b}_1, \ldots, \mathbf{b}_p$, then the product AB is the $m \times p$ matrix whose columns are $A\mathbf{b}_1, \ldots, A\mathbf{b}_p$. That is,

$$AB = A\begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_p \end{bmatrix} = \begin{bmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{b}_p \end{bmatrix}$$

Multiplication of matrices corresponds to composition of linear transformations.

Each column of AB is a linear combination of the columns of A using weights from the corresponding column of B.

Theorem 2. The following are the standard properties of matrix multiplication:

1. $A(BC) = (AB)C$
2. $A(B + C) = AB + AC$
3. $(B + C)A = BA + CA$
4. $r(AB) = (rA)B = A(rB)$ for any scalar r
5. $I_m A = A = AI_n$

In general, $AB \neq BA$; if $AB = AC$, then it is not true that $B = C$; and if AB is the zero matrix, then it can not be concluded that either $A = 0$ or $B = 0$.

Definition 5. If A is an $n \times n$ matrix and if k is a positive integer, then A^k denotes the product of k copies of A, i.e. $A^k = \underbrace{A \cdots A}_{k}$, where A^0 is interpreted as the identity matrix.

Definition 6. Given an $m \times n$ matrix A, the **transpose** of A is the $n \times m$ matrix, denoted by A^T, whose columns are formed from the corresponding rows of A.

Theorem 3. Let A and B denote matrices whose sizes are appropriate for the following sums and products.

1. $(A^T)^T = A$
2. $(A + B)^T = A^T + B^T$
3. $(rA)^T = rA^T$ for any scalar r
4. $(AB)^T = B^T A^T$

1.2.2 The Inverse of a Matrix

Definition 7. An $n \times n$ matrix A is said to be **invertible** if there is an $n \times n$ matrix C such that $AC = I_n = CA$. In this case, C is the unique inverse of A.

Definition 8. A matrix that is not invertible is called a **singular matrix** and an invertible matrix is also called a **nonsingular matrix**.

Theorem 4. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. If $ad - bc \neq 0$, then A is invertible and

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

If $ad - bc = 0$, then A is not invertible.

Definition 9. The quantity $\det A = ad - bc$ is called the **determinant** of 2×2 matrix A.

Theorem 5. If A is an invertible $n \times n$ matrix, then for each $\mathbf{b}$ in $\mathbb{R}^n$, the equation $A\mathbf{x} = \mathbf{b}$ has the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.

Theorem 6. The following are facts about invertible matrices:

1. If A is an invertible matrix, then A^{-1} is invertible and $(A^{-1})^{-1} = A$
2. If A and B are $n \times n$ invertible matrices, then so is AB and the inverse of AB is the product of the inverses of A and B in the reverse order. That is, $(AB)^{-1} = B^{-1}A^{-1}$
3. If A is an invertible matrix, then so is A^T and the inverse of A^T is the transpose of A^{-1}. That is, $(A^T)^{-1} = (A^{-1})^T$

Definition 10. An **elementary matrix** is one that is obtained by performing a single elementary row operation on an identity matrix.

If an elementary row operation is performed on an $m \times n$ matrix A, the resulting matrix can be written as EA, where the $m \times m$ matrix E is created by performing the same row operation on I_m.

Each elementary matrix E is invertible. The inverse of E is the elementary matrix of the same type that transforms E back into I.

Theorem 7. An $n \times n$ matrix A is invertible if and only if A is row equivalent to I_n, and in this case, any sequence of elementary row operations that reduces A to I_n also transforms I_n into A^{-1}.

The following is an algorithm for finding A^{-1}:

1. Place A and I side-by-side to form the augmented matrix $\begin{bmatrix} A & I \end{bmatrix}$.
2. Row reduce the augmented matrix.
3. If A is row equivalent to I, then $\begin{bmatrix} A & I \end{bmatrix}$ is row equivalent to $\begin{bmatrix} I & A^{-1} \end{bmatrix}$. Otherwise, A does not have an inverse.

1.2.3 Characterizations of Invertible Matrices

Theorem 8. (The Invertible Matrix Theorem) Let A be a square $n \times n$ matrix. Then the following statements are equivalent:

1. A is an invertible matrix.
2. A is row equivalent to the $n \times n$ identity matrix.
3. A has n pivot positions.
4. The equation $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
5. The columns of A form a linearly independent set.
6. The linear transformation $x \mapsto A\mathbf{x}$ is one-to-one.
7. The equation $A\mathbf{x} = \mathbf{b}$ has at least one solution for each $\mathbf{b}$ in $\mathbb{R}^n$.
8. The columns of A span $\mathbb{R}^n$.
9. The linear transformation $x \mapsto A\mathbf{x}$ maps $\mathbb{R}^n$ onto $\mathbb{R}^n$.
10. There is an $n \times n$ matrix C such that $CA = I$.
11. There is an $n \times n$ matrix D such that $AD = I$.
12. A^T is an invertible matrix.

Note that the Invertible Matrix Theorem applies only to square matrices.

Let A and B be square matrices. If $AB = I$, then A and B are both invertible, with $B = A^{-1}$ and $A = B^{-1}$.

Definition 11. A linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is **invertible** if there exists a function $S : \mathbb{R}^n \to \mathbb{R}^n$ such that $S(T(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$ and $T(S(\mathbf{x})) = \mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

Theorem 9. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation and let A be the standard matrix for T. Then T is invertible if and only if A is an invertible matrix. In that case, the linear transformation S given by $S(\mathbf{x}) = A^{-1}\mathbf{x}$ is unique.

1.2.4 Matrix Factorizations

Definition 12. A **factorization** of a matrix A is an equation that expresses A as a product of two or more matrices.

The LU factorization is motivated by the problem of solving a sequence of equations which all have the same coefficient matrix, i.e. $A\mathbf{x} = \mathbf{b}_1, \ldots, A\mathbf{x} = \mathbf{b}_p$. When A is invertible, it is costly to compute A^{-1} and then compute $A^{-1}\mathbf{b}_1, \ldots, A^{-1}\mathbf{b}_p$. It is more efficient to solve the first equation in the sequence by row reduction and obtain an LU factorization at the same time.

Definition 13. Assume that A is an $m \times n$ matrix that can be row reduced to echelon form without row interchanges. Then A can be written in the form $A = LU$, where L is an $m \times m$ lower triangular matrix with 1s on the diagonal and U is an $m \times n$ echelon form of A. Such a factorization is called an **LU factorization** of A. The matrix L is invertible and is called a unit lower triangular matrix. The following is an LU factorization:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ * & 1 & 0 & 0 \\ * & * & 1 & 0 \\ * & * & * & 1 \end{bmatrix} \begin{bmatrix} \blacksquare & * & * & * & * \\ 0 & \blacksquare & * & * & * \\ 0 & 0 & 0 & \blacksquare & * \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

When $A = LU$, the equation $A\mathbf{x} = \mathbf{b}$ can be written as $L(U\mathbf{x}) = \mathbf{b}$. We can thus find $\mathbf{x}$ by solving the pair of equations $L\mathbf{y} = b$, $U\mathbf{x} = y$. First, solve $L\mathbf{y} = \mathbf{b}$ for $\mathbf{y}$ and then solve $U\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$. Each equation is easy to solve because L and U are triangular.

The following is an algorithm for LU factorization:

1. Reduce A to an echelon form U by a sequence of row replacement operations, if possible.
2. Place entries in L such that the same sequence of row operations reduces L to I.

When the first step is possible, an LU factorization exists.

1.3 Determinants

1.3.1 Introduction to Determinants

Consider matrix A with $a_{11} \neq 0$. If we multiply the second and third rows of A by a_{11} and subtract appropriate multiples of the first row from the other two rows, we find that A is row equivalent to the following two matrices:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{11}a_{21} & a_{11}a_{22} & a_{11}a_{23} \\ a_{11}a_{31} & a_{11}a_{32} & a_{11}a_{33} \end{bmatrix} \sim \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{11}a_{22} - a_{12}a_{21} & a_{11}a_{23} - a_{13}a_{21} \\ 0 & a_{11}a_{32} - a_{12}a_{31} & a_{11}a_{33} - a_{13}a_{31} \end{bmatrix}$$

Since A is invertible, either the $(2,2)$-entry or the $(3,2)$-entry in the right matrix is nonzero. Suppose that the $(2,2)$-entry is nonzero. Multiply row 3 by $a_{11}a_{22} - a_{12}a_{21}$ and then to the new row 3, add $-(a_{11}a_{32} - a_{12}a_{31})$ times row 2. This will show that

$$A \sim \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{11}a_{22} - a_{12}a_{21} & a_{11}a_{23} - a_{13}a_{21} \\ 0 & 0 & a_{11}\Delta \end{bmatrix}$$

where

$$\begin{aligned} \Delta &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \\ &= a_{11} \cdot \det \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} - a_{12} \cdot \det \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} + a_{13} \cdot \det \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \\ &= a_{11} \cdot \det A_{11} - a_{12} \cdot \det A_{12} + a_{13} \cdot \det A_{13} \end{aligned}$$

and A_{11}, A_{12}, A_{13} are obtained from A by deleting the first row and one of the three columns. For any square matrix, we let A_{ij} denote the submatrix formed by deleting the i^{th} row and j^{th} column of A.

We call Δ the determinant of the 3×3 matrix A. Since A is invertible, Δ must be nonzero.

Definition 1. For $n \geq 2$, the **determinant** of an $n \times n$ matrix A is the sum of n terms of the form $\pm a_{1j} \det A_{1j}$, with plus and minus signs alternating, where the entries $a_{11}, \ldots, a_{1n}$ are from the first row of A, i.e.

$$\begin{aligned} \det A &= a_{11} \cdot \det A_{11} - a_{12} \cdot \det A_{12} + \cdots + (-1)^{1+n} a_{1n} \cdot \det A_{1n} \\ &= \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det A_{1j} \end{aligned}$$

Definition 2. Given matrix A, the $(\mathbf{i},\mathbf{j})$**-cofactor** of A is the number C_{ij} given by

$$C_{ij} = (-1)^{i+j} \det A_{ij}$$

Theorem 1. The determinant of an $n \times n$ matrix A can be computed by a cofactor expansion across any row or down any column. The expansion across the i^{th} row is

$$\det A = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in}$$

The cofactor expansion down the j^{th} column is

$$\det A = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj}$$

Theorem 2. If A is a triangular matrix, then $\det A$ is the product of the entries on the main diagonal of A.

1.3.2 Properties of Determinants

Theorem 3. Let A be a square matrix.

1. If a multiple of one row of A is added to another row to produce a matrix B, then $\det B = \det A$.
2. If two rows of A are interchanged to produce B, then $\det B = -\det A$.
3. If one row of A is multiplied by k to produce B, then $\det B = k \cdot \det A$.

Suppose a square matrix A has been reduced to an echelon form U by row replacements and row interchanges. If there are r interchanges, then $\det A = (-1)^r \det U$. Since U is in echelon form, it is triangular, and so $\det U$ is the product of all the diagonal entries $u_{11}, \ldots, u_{nn}$. If A is invertible, the entries u_{ii} are all pivots. Otherwise, at least u_{nn} is zero, and the product $u_{11} \cdots u_{nn}$ is zero. Thus,

$$\det A = \begin{cases} (-1)^r \cdot (\text{product of pivots in } U) & \text{when } A \text{ is invertible} \\ 0 & \text{when } A \text{ is not invertible} \end{cases}$$

We note that although the echelon form U is not unique and the pivots are not unique, the product of the pivots is unique, except for a possible minus sign.

Theorem 4. A square matrix A is invertible if and only if $\det A \neq 0$.

We can perform operations on the columns of a matrix in a way that is analogous to the row operations we have considered. Therefore, we have the following theorem.

Theorem 5. If A is an $n \times n$ matrix, then $\det A^T = \det A$.

Theorem 6. If A and B are $n \times n$ matrices, then $\det AB = (\det A)(\det B)$.

If all columns except one are held fixed, then $\det A$ is a linear function of that one vector variable.

1.3.3 Cramer's Rule

For any $n \times n$ matrix A and any $\mathbf{b}$ in $\mathbb{R}^n$, let $A_i(b)$ be the matrix obtained from A by replacing the column i by the vector $\mathbf{b}$.

Theorem 7. (Cramer's Rule) Let A be an invertible $n \times n$ matrix. For any $\mathbf{b}$ in $\mathbb{R}^n$, the unique solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$ has entries given by

$$x_i = \frac{\det A_i(\mathbf{b})}{\det A}$$

for $i = 1, 2, \ldots, n$.

Cramer's Rule leads easily to a general formula for the inverse of an $n \times n$ matrix A. The j^{th} column of A^{-1} is a vector $\mathbf{x}$ that satisfies $A\mathbf{x} = \mathbf{e}_j$ and the i^{th} entry of $\mathbf{x}$ is the (i, j)-entry of A^{-1}. By Cramer's rule, $A^{-1}_{ij} = \frac{\det A_i(\mathbf{e}_j)}{\det A}$. We recall that a cofactor expansion down column i of $A_i(\mathbf{e}_j)$ shows that $\det A_i(\mathbf{e}_j) = (-1)^{i+j} \det A_j i = C_{ji}$ and hence

$$A^{-1} = \frac{1}{\det A} \begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix}$$

Definition 3. The matrix on the right side is called the **adjugate** of A, denoted by adj A.

Theorem 8. Let A be an invertible $n \times n$ matrix. Then $A^{-1} = \frac{1}{\det A} \text{adj} A$.

Theorem 9. If A is a 2×2 matrix, then the area of the parallelogram determined by the columns of A is $|\det A|$. If A is a 3×3 matrix, the volume of the parallelepiped determined by the columns of A is $|\det A|$.

Determinants can be used to describe an important property of linear transformations in $\mathbb{R}^2$ and $\mathbb{R}^3$ as well.

Theorem 10. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation determined by a 2×2 matrix A. If S is a parallelogram in $\mathbb{R}^2$, then

$$\{\text{area of } T(S)\} = |\det A| \cdot \{\text{area of } S\}$$

If T is determined by a 3×3 matrix A and if S is a parallelepiped in $\mathbb{R}^3$, then

$$\{\text{volume of } T(S)\} = |\det A| \cdot \{\text{volume of } S\}$$

1.4 Vector Spaces

1.4.1 Vector Spaces and Subspaces

Definition 1. A **vector space** is a nonempty set V of objects, called vectors, on which two operations are defined, called addition and multiplication by scalars, subject to the ten axioms below. The axioms must hold for all vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in V and for all scalars c and d.

1. The sum of $\mathbf{u}$ and $\mathbf{v}$, denoted by $\mathbf{u}+\mathbf{v}$, is in V.
2. $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$.
3. $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$.
4. There is a zero vector $\mathbf{0}$ in V such that $\mathbf{u}+\mathbf{0}=\mathbf{u}$.
5. For each $\mathbf{u}$ in V, there is a vector $-\mathbf{u}$ in V such that $\mathbf{u}+(-\mathbf{u})=\mathbf{0}$.
6. The scalar multiple of $\mathbf{u}$ by c, denoted by $c\mathbf{u}$, is in V.
7. $c(\mathbf{u}+\mathbf{v})=c\mathbf{u}+c\mathbf{v}$.
8. $(c+d)\mathbf{u}=c\mathbf{u}+d\mathbf{u}$.
9. $c(d\mathbf{u})=(cd)\mathbf{u}$.
10. $1\mathbf{u}=\mathbf{u}$.

Using only these axioms, one can show that the zero vector, $\mathbf{0}$, is unique and the vector $-\mathbf{u}$, called the **negative** of $\mathbf{u}$ is unique for each $\mathbf{u}$ in V.

For each $\mathbf{u}$ in V and scalar c, $0\mathbf{u}=\mathbf{0}$, $c\mathbf{0}=\mathbf{0}$, and $-\mathbf{u}=(-1)\mathbf{u}$.

Definition 2. A **subspace** of a vector space V is a subset H of V that has three properties:

1. The zero vector of V is in H.
2. For each $\mathbf{u}$ and $\mathbf{v}$ in H, the sum $\mathbf{u}+\mathbf{v}$ is in H.
3. For each $\mathbf{u}$ in H and each scalar c, the vector $c\mathbf{u}$ is in H.

i.e. a subspace is closed under addition and scalar multiplication.

Every subspace is a vector space. Conversely, every vector space is a subspace (of itself and possibly of larger spaces). The term subspace is used when at least two vector spaces are in mind, with one inside the other, and the phrase **subspace of** V identifies V as the larger space.

The set consisting of only the zero vector in a vector space V is a subspace of V, called the **zero subspace** and written as $\{\mathbf{0}\}$.

Let $\mathbb{P}$ be the set of all polynomials with real coefficients, with operations in $\mathbb{P}$ defined as for functions. Then $\mathbb{P}$ is a subspace of the space of all real-valued functions defined on $\mathbb{R}$. Also, for each $n \geq 0$, $\mathbb{P}_n$ is a subspace of $\mathbb{P}$, because $\mathbb{P}_n$ is a subset of $\mathbb{P}$ that contains the zero polynomial, the sum of two polynomials in $\mathbb{P}_n$ is also in $\mathbb{P}_n$, and a scalar multiple of a polynomial in $\mathbb{P}_n$ is also in $\mathbb{P}_n$.

Theorem 1. If $\mathbf{v}_1, \ldots, \mathbf{v}_p$ are in a vector space V, then $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is a subspace of V.

Definition 3. We call $\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ **the subspace spanned** by $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$. Given any subspace H of V, a **spanning set** for H is a set $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ in H such that $H = \text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$.

1.4.2 Null Spaces, Column Spaces, and Linear Transformations

Definition 4. The **null space** of an $m \times n$ matrix A, written as Nul A, is the set of all solutions of the homogeneous equation $A\mathbf{x} = \mathbf{0}$, i.e.

$$\text{Nul } A = \{\mathbf{x} : \mathbf{x} \text{ is in } \mathbb{R}^n \text{ and } A\mathbf{x} = \mathbf{0}\}$$

Theorem 2. The null space of an $m \times n$ matrix A is a subspace of $\mathbb{R}^n$. Equivalently, the set of all solutions to a system of $A\mathbf{x} = \mathbf{0}$ of m homogeneous linear equations in n unknowns is a subspace of $\mathbb{R}^n$.

Definition 5. The **column space** of an $m \times n$ matrix A, written as Col A, is the set of all linear combinations of the columns of A. If $A = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{bmatrix}$, then

$$\text{Col } A = \text{ Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$$

Theorem 3. The column space of an $m \times n$ matrix A is a subspace of $\mathbb{R}^m$.

We note that a typical vector in Col A can be written as $A\mathbf{x}$ for some $\mathbf{x}$ because the notation $A\mathbf{x}$ stands for a linear combination of the columns of A. That is,

$$\text{Col } A = \{\mathbf{b} : \mathbf{b} = A\mathbf{x} \text{ for some } \mathbf{x} \text{ in } \mathbb{R}^n\}$$

The column space of an $m \times n$ matrix A is all of $\mathbb{R}^m$ if and only if the equation $A\mathbf{x} = \mathbf{b}$ has a solution for each $\mathbf{b}$ in $\mathbb{R}^m$.

The table on the following page compares Nul A and Col A for an $m \times n$ matrix A.

Definition 6. A **linear transformation** T from a vector space V into a vector space W is a rule that assigns to each vector $\mathbf{x}$ in V a unique vector $T(\mathbf{x})$ in W such that

1. $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v}$ in V, and
2. $T(c\mathbf{u}) = cT(\mathbf{u})$ for all $\mathbf{u}$ in V and all scalars c.

Definition 7. The **kernel** (or **null space**) of such a T is the set of all $\mathbf{u}$ in V such that $T(\mathbf{u}) = \mathbf{0}$. The **range** of T is the set of all vectors in W of the form $T(\mathbf{x})$ for some $\mathbf{x}$ in V. If T happens to arise as a matrix transformation, $T(\mathbf{x}) = A\mathbf{x}$ for some matrix A, then the kernel and the range of T are simply the null space and the column space of A as defined earlier.

The kernel of T is a subspace of V and the range of T is a subspace of W.

Nul A is a subspace of $\mathbb{R}^n$.	Col A is a subspace of $\mathbb{R}^m$.
Nul A is implicitly defined; that is, you are given only a condition ($A\mathbf{x} = 0$) that vectors in Nul A must satisfy.	Col A is explicitly defined; that is, you are told how to build vectors in Col A.
It takes time to find vectors in Nul A. Row operations on $\begin{bmatrix} A & \mathbf{0} \end{bmatrix}$ are required.	It is easy to find vectors in Col A. The columns of A are displayed; others are formed from them.
There is no obvious relation between Nul A and the entries in A.	There is an obvious relation between Col A and the entries in A, since each column of A is in Col A.
A typical vector $\mathbf{v}$ in Nul A has the property that $A\mathbf{v} = 0$.	A typical vector $\mathbf{v}$ in Col A has the property that the equation $A\mathbf{x} = \mathbf{v}$ is consistent.
Given a specific vector $\mathbf{v}$, it is easy to tell if $\mathbf{v}$ is in Nul A. Just compute $A\mathbf{v}$.	Given a specific vector $\mathbf{v}$, it may take time to tell if $\mathbf{v}$ is in Col A. Row operations on $\begin{bmatrix} A & \mathbf{c} \end{bmatrix}$ are required.
Nul $A = \{\mathbf{0}\}$ if and only if the equation $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.	Col $A = \mathbb{R}^m$ if and only if the equation $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b}$ in $\mathbb{R}^m$.
Nul $A = \{\mathbf{0}\}$ if and only if the linear transformation $\mathbf{x} \mapsto A\mathbf{x}$ is one-to-one.	Col $A = \mathbb{R}^m$ if and only if the linear transformation $\mathbf{x} \mapsto A\mathbf{x}$ maps $\mathbb{R}^n$ to $\mathbb{R}^m$.

Figure 1.1: Comparison of Nul A and Col A

1.4.3 Linearly Independent Sets, Bases

Definition 8. An indexed set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ in V is said to be **linearly independent** if the vector equation

$$c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p$$

has only the trivial solution $c_1 = 0, \ldots, c_p = 0$.

An indexed set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is said to be **linearly dependent** if there is a non-trivial solution, that is, if there are some weights, $c_1, \ldots, c_p$, that are not all zero.

A set containing a single vector $\mathbf{v}$ is linearly independent if and only if $\mathbf{v} \neq \mathbf{0}$. A set of two vectors is linearly dependent if and only if one of the vectors is a multiple of the other. Any set containing the zero vector is linearly dependent.

Theorem 4. An indexed set $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ of two or more vectors is linearly dependent if and only if at least one of the vectors in S is a linear combination of the others. In fact, if S is linearly dependent and $\mathbf{v}_1 \neq \mathbf{0}$, then some $\mathbf{v}_j$ (with $j > 1$) is a linear combination of the preceding vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$.

The main difference between linear dependence in $\mathbb{R}^n$ and in a general vector space is that when the vectors are not n-tuples, the homogeneous equation usually cannot be written as a system of n linear equations. That is, the vectors cannot be made into the columns of a matrix A in order to study the equation $A\mathbf{x} = \mathbf{0}$.

Definition 9. Let H be a subspace of a vector space V. An indexed set of vectors $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ in V is a basis for H if

1. $\mathcal{B}$ is a linearly independent set, and
2. the subspace spanned by $\mathcal{B}$ coincides with H, that is $H = \text{Span}\{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$

Two ways to view a basis are as a spanning set that is as small as possible and a linearly independent set that is as large as possible.

The set $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is called the **standard basis** for $\mathbb{R}^n$ where

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ \cdots, \ \mathbf{e}_n = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

Theorem 5. (Spanning Set Theorem) Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ be a set in V and let $H = \text{Span}\{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$.

1. If one of the vectors in S, say $\mathbf{v}_k$, is a linear combination of the remaining vectors in S, then the set formed from S by removing $\mathbf{v}_k$ still spans H.
2. If $H \neq \{\mathbf{0}\}$, some subset of S is a basis for H.

Theorem 6. The pivot columns of a matrix A form a basis for Col A.

1.4.4 Coordinate Systems

Theorem 7. (Unique Representation Theorem) Let $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ be a basis for a vector space V. Then for each $\mathbf{x}$ in V, there exists a unique set of scalars $c_1, \ldots, c_n$ such that

$$\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$$

Definition 10. Suppose $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ is a basis for V and $\mathbf{x}$ is in V. The **coordinates of x relative to the basis $\mathcal{B}$** (or the **$\mathcal{B}$-coordinates of x**) are the weights $c_1, \ldots, c_n$ such that $\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$.

If $c_1, \ldots, c_n$ are the $\mathcal{B}$-coordinates of $\mathbf{b}$, then the vector in $\mathbb{R}^n$

$$[x]_{\mathcal{B}} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

is the **coordinate vector of x (relative to $\mathcal{B}$)**, or the **$\mathcal{B}$-coordinate vector of x**. The mapping $\mathbf{x} \mapsto [\mathbf{x}]_{\mathcal{B}}$ is the **coordinate mapping (determined by $\mathcal{B}$)**.

Definition 11. Let $P_{\mathcal{B}} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{bmatrix}$. Then the vector equation $\mathbf{x} = c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n$ is equivalent to $\mathbf{x} = P_{\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$. We call $P_{\mathcal{B}}$ the **change-of-coordinates matrix** from $\mathcal{B}$ to the standard basis in $\mathbb{R}^n$.

Since the columns of $P_{\mathcal{B}}$ form a basis for $\mathbb{R}^n$, $P_{\mathcal{B}}$ is invertible and left-multiplication by $P_{\mathcal{B}}^{-1}$ converts $\mathbf{x}$ into its $\mathcal{B}$-coordinate vector:

$$P_{\mathcal{B}}^{-1} = [\mathbf{x}]_{\mathcal{B}}$$

Theorem 8. Let $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ be a basis for a vector space V. Then the coordinate mapping $\mathbf{x} \mapsto [\mathbf{x}]_{\mathcal{B}}$ is a one-to-one linear transformation from V onto $\mathbb{R}^n$.

The coordinate mapping is a linear transformation and extends to linear combinations. If $\mathbf{u}_1, \ldots, \mathbf{u}_p$ are in V and if $c_1, \ldots, c_p$ are scalars, then

$$[c_1\mathbf{u}_1 + \cdots + c_p\mathbf{u}_p]_{\mathcal{B}} = c_1[\mathbf{u}_1]_{\mathcal{B}} + \cdots + c_p[\mathbf{u}_p]_{\mathcal{B}}$$

i.e. the $\mathcal{B}$-coordinate vector of a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_p$ is the same as the linear combination of their coordinate vectors.

The coordinate mapping above is an important example of an isomorphism from V onto $\mathbb{R}^n$.

Definition 12. A one-to-one linear transformation from a vector space V onto a vector space W is called an **isomorphism** from V onto W.

In general, if $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ is a basis for H, then the mapping $\mathbf{x} \mapsto [\mathbf{x}]_{\mathcal{B}}$ is an isomorphism.

The notation and terminology for V and W may differ, but the two spaces are indistinguishable as vector spaces. Every vector space calculation in V is accurately reproduced in W, and vice versa. In particular, any real vector space with a basis of n vectors is indistinguishable from $\mathbb{R}^n$.

1.4.5 The Dimension of a Vector Space

Theorem 9. If a vector space V has a basis $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$, then any set in V containing more than n vectors must be linearly dependent.

Theorem 10. If a vector space V has a basis of n vectors, then every basis of V must consist of exactly n vectors.

If a nonzero vector space V is spanned by a finite set S, then a subset of S is a basis for V.

Definition 13. If V is spanned by a finite set, then V is said to be **finite-dimensional** and the **dimension** of V, written as $\dim V$, is the number of vectors in a basis for V. The dimension of the zero vector space $\{\mathbf{0}\}$ is defined to be zero. If V is not spanned by a finite set, then V is said to be **infinite-dimensional**.

The space $\mathbb{R}^n$ has dimension n. Every basis for $\mathbb{R}^n$ consists of n vectors.

Theorem 11. Let H be a subspace of a finite-dimensional vector space V. Any linearly independent set in H can be expanded, if necessary, to a basis for H. Also, H is finite-dimensional and

$$\dim H \leq \dim V$$

Theorem 12. (Basis Theorem) Let V be a p-dimensional vector space, $p \geq 1$. Any linearly independent set of exactly p elements in V is automatically a basis for V. Any set of exactly p elements that spans V is automatically a basis for V.

The dimension of Nul A is the number of free variables in the equation $A\mathbf{x} = \mathbf{0}$. The dimension of Col A is the number of pivot columns in A.

1.4.6 Rank

Definition 14. If A is an $m \times n$ matrix, each row of A has n entries and thus can be identified with a vector in $\mathbb{R}^n$. The set of all linear combinations of the row vectors is called the **row space** of A and is denoted Row A.

Each row has n entries, so Row A is a subspace of $\mathbb{R}^n$. Since the rows of A are identified with the columns of A^T, we could also write Col A^T in place of Row A.

Theorem 13. If two matrices A and B are row equivalent, then their row spaces are the same. If B is in echelon form, the nonzero rows of B form a basis for the row space of A as well as for that of B.

Definition 15. The **rank** of A is the dimension of the column space of A.

Since Row A is the same as Col A^T, the dimension of the row space of A is the rank of A^T. The dimension of the null space is sometimes called the **nullity** of A.

Theorem 14. (Rank Theorem) The dimensions of the column space and the row space of an $m \times n$ matrix A are equal. This common dimension, the rank of A, also equals the number of pivot positions in A and satisfies the equation

$$\text{rank } A + \dim \text{ Nul } A = n$$

Theorem 15. (The Invertible Matrix Theorem continued) Let A be a square $n \times n$ matrix. Then the following statements are equivalent:

13. The columns of A form a basis of $\mathbb{R}^n$.
14. Col $A = \mathbb{R}^n$
15. dim Col $A = n$
16. rank $A = n$
17. Nul $A = \{\mathbf{0}\}$
18. dim Nul $A = 0$

1.4.7 Change of Basis

Theorem 16. Let $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ and $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ be bases of a vector space V. Then there is a unique $n \times n$ matrix $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}$ such that

$$[\mathbf{x}]_{\mathcal{C}} = \underset{\mathcal{C}\leftarrow\mathcal{B}}{P}[\mathbf{x}]_{\mathcal{B}}$$

The columns of $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}$ are the $\mathcal{C}$-coordinate vectors of the vectors in the basis $\mathcal{B}$, i.e.

$$\underset{\mathcal{C}\leftarrow\mathcal{B}}{P} = \begin{bmatrix}[\mathbf{b}_1]_{\mathcal{C}} & \cdots & [\mathbf{b}_n]_{\mathcal{C}}\end{bmatrix}$$

Definition 16. The matrix $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}$ is called the **change-of-coordinates matrix from $\mathcal{B}$ to $\mathcal{C}$**. Multiplication by $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}$ changes $\mathcal{B}$-coordinates into $\mathcal{C}$-coordinates.

The columns of $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}$ are linearly independent because they are the coordinate vectors of the linearly independent set $\mathcal{B}$. Since $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}$ is square, it must be invertible, and left-multiplying by $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}^{-1}$ yields

$$\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}^{-1}[\mathbf{x}]_{\mathcal{C}} = [\mathbf{x}]_{\mathcal{B}}$$

Thus, $\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}^{-1}$ is the matrix that converts $\mathcal{C}$-coordinates into $\mathcal{B}$-coordinates, i.e.

$$\underset{\mathcal{C}\leftarrow\mathcal{B}}{P}^{-1} = \underset{\mathcal{B}\leftarrow\mathcal{C}}{P}$$

If $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ and $\mathcal{E}$ is the standard basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ in $\mathbb{R}^n$, then $[\mathbf{b}_1]_{\mathcal{E}} = \mathbf{b}_1$ and likewise for the other vectors in $\mathcal{B}$. In this case, $\underset{\mathcal{E}\leftarrow\mathcal{B}}{P}$ is the same as the change-of-coordinates matrix $P_{\mathcal{B}}$, namely $P_{\mathcal{B}} = \begin{bmatrix}\mathbf{b}_1 & \cdots & \mathbf{b}_n\end{bmatrix}$.

In general, we can use a row reduction procedure to find the change-of-coordinates matrix between any two bases in $\mathbb{R}^n$, i.e.

$$\begin{bmatrix}\mathbf{c}_1 & \cdots & \mathbf{c}_n & | & \mathbf{b}_1 & \cdots & \mathbf{b}_n\end{bmatrix} \sim \left[\ I \mid \underset{\mathcal{C}\leftarrow\mathcal{B}}{P}\ \right]$$

1.5 Eigenvalues and Eigenvectors

1.5.1 Eigenvectors and Eigenvalues

Definition 1. An **eigenvector** of an $n \times n$ matrix A is a nonzero vector $\mathbf{x}$ such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar λ. A scalar λ is called an **eigenvalue** of A if there is a nontrivial solution $\mathbf{x}$ of $A\mathbf{x} = \lambda\mathbf{x}$; such an $\mathbf{x}$ is called an **eigenvector corresponding to** λ.

The value λ is an eigenvalue of an $n \times n$ matrix A if and only if the equation

$$(A - \lambda I)\mathbf{x} = \mathbf{0}$$

has a nontrivial solution.

Definition 2. The set of all solutions of $(A-\lambda I)\mathbf{x} = \mathbf{0}$ is the null space of the matrix $A-\lambda I$ and is called the **eigenspace** of A corresponding to λ. The eigenspace of A is a subspace of $\mathbb{R}^n$.

Theorem 1. The eigenvalues of a triangular matrix are the entries on its main diagonal.

The value 0 is an eigenvalue if and only if A not invertible.

Theorem 2. If $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are eigenvectors that correspond to distinct eigenvalues $\lambda_1, \ldots, \lambda_r$ of an $n \times n$ matrix A, then the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ is linearly independent.

1.5.2 The Characteristic Equation

To find the eigenvalues of $A = \begin{bmatrix} 2 & 3 \\ 3 & -6 \end{bmatrix}$, we must find all scalars λ such that the matrix equation $(A - \lambda I)\mathbf{x} = \mathbf{0}$ has a nontrivial solution. This is equivalent to finding all λ such that the matrix $A - \lambda I$ is not invertible, where

$$A - \lambda I = \begin{bmatrix} 2 & 3 \\ 3 & -6 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 2-\lambda & 3 \\ 3 & -6-\lambda \end{bmatrix}$$

This matrix fails to be invertible precisely when the determinant is zero, so the eigenvalues of A are the solutions of the equation

$$\det(A - \lambda I) = \det \begin{bmatrix} 2-\lambda & 3 \\ 3 & -6-\lambda \end{bmatrix} = 0$$

Recalling the formula for the determinant of a 2×2 matrix, we have that

$$\begin{aligned} \det(A - \lambda I) &= (2-\lambda)(-6-\lambda) - (3)(3) \\ &= -12 + 6\lambda - 2\lambda + \lambda^2 - 9 \\ &= \lambda^2 + 4\lambda - 21 \\ &= (\lambda - 3)(\lambda + 7) \end{aligned}$$

If $\det(A - \lambda I) = 0$, then $\lambda = 3$ or $\lambda = -7$, so the eigenvalues of A are 3 and -7.

We can add to the Invertible Matrix Theorem:

Theorem 3. (Invertible Matrix Theorem continued) Let A be an $n \times n$ matrix. Then A is invertible if and only if:

19. The number 0 is not an eigenvalue of A.
20. The determinant of A is not zero.

Definition 3. The equation $\det(A - \lambda I) = 0$ is called the **characteristic equation** of A.

A scalar λ is an eigenvalue of an $n \times n$ matrix A if and only if λ satisfies the characteristic equation

$$\det(A - \lambda I) = 0$$

Definition 4. If A is an $n \times n$ matrix, then $\det(A - \lambda I)$ is a polynomial of degree n called the **characteristic polynomial** of A.

Definition 5. The **(algebraic) multiplicity** of an eigenvalue λ is its multiplicity as a root of the characteristic equation.

Definition 6. If A and B are $n \times n$ matrices, then A **is similar to** B if there is an invertible matrix P such that $P^{-1}AP = B$ or equivalently $A = PBP^{-1}$.

Writing Q for P^{-1}, we have $Q^{-1}BQ = A$, so B is also similar to A, and we can simply say that A and B are **similar**. Changing A into $P^{-1}AP$ is called a similarity transformation.

Theorem 4. If $n \times n$ matrices A and B are similar, then they have the same characteristic polynomial and hence the same eigenvalues (with the same multiplicities).

1.5.3 Diagonalization

In many cases, the eigenvalue-eigenvector information contained within a matrix A can be displayed in a useful factorization of the form $A = PDP^{-1}$ where D is a diagonal matrix.

Definition 7. A square matrix A is said to be **diagonalizable** if A is similar to a diagonal matrix, i.e. $A = PDP^{-1}$ for some invertible matrix P and some diagonal matrix D.

Theorem 5. An $n \times n$ matrix A is diagonalizable if and only if A has n linearly independent eigenvectors. In fact, $A = PDP^{-1}$, with D a diagonal matrix, if and only if the columns of P are n linearly independent eigenvectors of A. In this case, the diagonal entries of D are eigenvalues of A that correspond, respectively, to the eigenvectors in P.

A is diagonalizable if and only if there are enough eigenvectors to form a basis of $\mathbb{R}^n$. We call such a basis an **eigenvector basis** of $\mathbb{R}^n$.

Theorem 6. An $n \times n$ matrix with n distinct eigenvalues is diagonalizable.

The above is a sufficient but not necessary condition for a matrix to be diagonalizable.

When A is diagonalizable but has fewer than n distinct eigenvalues, it is still possible to build P in a way that makes P automatically invertible.

Theorem 7. Let A be an $n \times n$ matrix whose distinct eigenvalues are $\lambda_1, \ldots, \lambda_p$.

1. For $1 \leq k \leq p$, the dimension of the eigenspace for λ_k is less than or equal to the multiplicity of the eigenvalue λ_k.
2. The matrix A is diagonalizable if and only if the sum of the dimensions of the eigenspaces equals n, and this happens if and only if (i) the characteristic polynomial factors completely into linear factors and (ii) the dimension of the eigenspace for each λ_k equals the multiplicity of λ_k.
3. If A is diagonalizable and $\mathcal{B}_k$ is a basis for the eigenspace corresponding to λ_k for each k, then the total collection of vectors in the sets $\mathcal{B}_1, \ldots, \mathcal{B}_p$ forms an eigenvector basis for $\mathbb{R}^n$.

Having a factorization of the form $A = PDP^{-1}$ where D is a diagonal matrix enables us to compute A^k quickly for large values of k.

The powers of a diagonal matrix are easy to compute, as seen below.

If $D = \begin{bmatrix} 5 & 0 \\ 0 & 3 \end{bmatrix}$, then $D^2 = \begin{bmatrix} 5 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 5^2 & 0 \\ 0 & 3^2 \end{bmatrix}$. In general, $D^k = \begin{bmatrix} 5^k & 0 \\ 0 & 3^k \end{bmatrix}$ for $k \geq 1$

When $A = PDP^{-1}$ for some invertible P and diagonal D, A^k is also easy to compute, as seen below.

If $A = \begin{bmatrix} 7 & 2 \\ -4 & 1 \end{bmatrix}$, then $A = PDP^{-1}$ where $P = \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix}$ and $D = \begin{bmatrix} 5 & 0 \\ 0 & 3 \end{bmatrix}$.

By associativity of matrix multiplication,

$$\begin{aligned} A^2 &= (PDP^{-1})(PDP^{-1}) = PD(P^{-1}P)DP^{-1} = PD^2P^{-1} \\ &= \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 5^2 & 0 \\ 0 & 3^2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix} \end{aligned}$$

In general,

$$\begin{aligned} A^k = PD^kP^{-1} &= \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 5^k & 0 \\ 0 & 3^k \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 2 \cdot 5^k - 3^k & 5^k - 3^k \\ 2 \cdot 3^k - 2 \cdot 5^k & 2 \cdot 3^k - 5^k \end{bmatrix} \end{aligned}$$

1.5.4 Eigenvectors and Linear Transformations

Let V be an n-dimensional vector space, let W be an m-dimensional vector space, and let T be any linear transformation from V to W. To associate a matrix with T, choose (ordered) bases $\mathcal{B}$ and $\mathcal{C}$ for V and W, respectively. Given any $\mathbf{x}$ in V, the coordinate vector $[\mathbf{x}]_{\mathcal{B}}$ is in $\mathbb{R}^n$ and the coordinate vector of its image, $[T(\mathbf{x})]_{\mathcal{C}}$, is in $\mathbb{R}^m$.

Let $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ be the basis $\mathcal{B}$ for V. If $\mathbf{x} = r_1\mathbf{b}_1 + \cdots + r_n\mathbf{b}_n$, then

$$[\mathbf{x}]_{\mathcal{B}} = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}$$

and

$$T(\mathbf{x}) = T(r_1\mathbf{b}_1 + \cdots + r_n\mathbf{b}_n) = r_1T(\mathbf{b}_1) + \cdots + r_nT(\mathbf{b}_n)$$

because T is linear. This leads to

$$[T(\mathbf{x})]_{\mathcal{C}} = r_1[T(\mathbf{b}_1)]_{\mathcal{C}} + \cdots + r_n[T(\mathbf{b}_n)]_{\mathcal{C}}$$

Since $\mathcal{C}$-coordinate vectors are in $\mathbb{R}^m$, the vector equation above can be written as a matrix equation, namely

$$[T(\mathbf{x})]_{\mathcal{C}} = M[\mathbf{x}]_{\mathcal{B}}$$

where

$$M = \begin{bmatrix} [T(\mathbf{b}_1)]_{\mathcal{C}} & \cdots & [T(\mathbf{b}_n)]_{\mathcal{C}} \end{bmatrix}$$

Definition 8. The matrix M is a matrix representation of T, called the **matrix for T relative to the bases $\mathcal{B}$ and $\mathcal{C}$**.

If $\mathcal{B}$ and $\mathcal{C}$ are bases for the same space V and if T is the identity transformation $T(\mathbf{x}) = \mathbf{x}$ for $\mathbf{x}$ in V, then matrix M is simply a change-of-coordinates matrix.

Definition 9. In the common case where W is the same as V and the basis $\mathcal{C}$ is the same as $\mathcal{B}$, the matrix M is called the **matrix for T relative to $\mathcal{B}$** or simply the **$\mathcal{B}$-matrix for** T and is denoted by $[T]_{\mathcal{B}}$.

The $\mathcal{B}$-matrix for $T : V \to V$ satisfies $[T(\mathbf{x})]_{\mathcal{B}} = [T]_{\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$ for all $\mathbf{x}$ in V.

Theorem 8. (Diagonal Matrix Representation) Suppose $A = PDP^{-1}$, where D is a diagonal $n \times n$ matrix. If $\mathcal{B}$ is the basis for $\mathbb{R}^n$ formed from the columns of P, then D is the $\mathcal{B}$-matrix for the transformation $\mathbf{x} \mapsto A\mathbf{x}$.

The set of all matrices similar to a matrix A coincides with the set of all matrix representations of the transformation $\mathbf{x} \mapsto A\mathbf{x}$.

1.5.5 Complex Eigenvalues

Since the characteristic equation of an $n \times n$ matrix involves a polynomial of degree n, the equation always has exactly n roots, counting multiplicities, provided that possibly complex roots are included.

If the characteristic equation of a real matrix A has some complex roots, then these roots provide critical information about A. The key is to let A act on the space $\mathbb{C}^n$ of n-tuples of complex numbers.

The matrix eigenvalue-eigenvector theory developed for $\mathbb{R}^n$ applies equally well to $\mathbb{C}^n$.

Definition 10. A complex scalar λ satisfies $\det(A - \lambda I) = 0$ if and only if there is a nonzero vector $\mathbf{x}$ in $\mathbb{C}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$. We call λ a **(complex) eigenvalue** and $\mathbf{x}$ a **(complex) eigenvector** corresponding to λ.

When A is real, its complex eigenvalues occur in conjugate pairs.

Theorem 9. Let A be a real 2×2 matrix with a complex eigenvalue $\lambda = a - bi$ $(b \neq 0)$ and an associated eigenvector $\mathbf{v}$ in $\mathbb{C}^2$. Then

$$A = PCP^{-1}$$

where $P = \begin{bmatrix} \text{Re } \mathbf{v} & \text{Im } \mathbf{v} \end{bmatrix}$ and $C = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$.

1.6 Orthogonality and Least Squares

1.6.1 Inner Product, Length, Orthogonality

If $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^n$, then we regard $\mathbf{u}$ and $\mathbf{v}$ as $n \times 1$ matrices. The transpose $\mathbf{u}^T$ is a $1 \times n$ matrix, and the matrix product $\mathbf{u}^T\mathbf{v}$ is a 1×1 matrix, which we write as a scalar.

Definition 1. The number $\mathbf{u}^T\mathbf{v}$ is called the **inner product** of $\mathbf{u}$ and $\mathbf{v}$ and is often written as $\mathbf{u} \cdot \mathbf{v}$. This inner product is also referred to as a **dot product**.

For vectors $\mathbf{u}, \mathbf{v}$ with n entries, the inner product of $\mathbf{u}$ and $\mathbf{v}$ is

$$\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = u_1v_1 + u_2v_2 + \cdots + u_nv_n$$

Theorem 1. Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be vectors in $\mathbb{R}^n$ and let c be a scalar. Then

1. $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$
2. $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$
3. $(c\mathbf{u}) \cdot \mathbf{v} = c(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \cdot (c\mathbf{v})$
4. $\mathbf{u} \cdot \mathbf{u} \geq 0$ and $\mathbf{u} \cdot \mathbf{u} = 0$ if and only if $\mathbf{u} = 0$

We can combine the second and third properties to obtain

$$(c_1\mathbf{u}_1 + \cdots + c_p\mathbf{u}_p) \cdot \mathbf{w} = c_1(\mathbf{u}_1 \cdot \mathbf{w}) + \cdots + c_p(\mathbf{u}_p \cdot \mathbf{w})$$

If $\mathbf{v}$ is in $\mathbb{R}^n$, with entries $v_1, \ldots, v_n$, then the square root of $\mathbf{v} \cdot \mathbf{v}$ is defined because $\mathbf{v} \cdot \mathbf{v}$ is nonnegative.

Definition 2. The **length** (or **norm**) of $\mathbf{v}$ is the nonnegative scalar $\|\mathbf{v}\|$ defined by

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \text{ and } \|\mathbf{v}\|^2 = \mathbf{v} \cdot \mathbf{v}$$

Suppose $\mathbf{v}$ is in $\mathbb{R}^2$. If we identify $\mathbf{v}$ with a geometric point in the plane, as usual, then $\|\mathbf{v}\|$ coincides with the standard notion of the length of the line segment from the origin to $\mathbf{v}$. The same holds for $\mathbf{v}$ in $\mathbb{R}^3$.

For any scalar, the length of $c\mathbf{v}$ is $|c|$ times the length of $\mathbf{v}$, i.e.

$$\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$$

Definition 3. A vector whose length is 1 is called a **unit vector**. If we divide a nonzero vector $\mathbf{v}$ by its length, we obtain a unit vector $\mathbf{u}$, a process called **normalization**.

Definition 4. For $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, the **distance between u and v**, written as $\text{dist}(\mathbf{u}, \mathbf{v})$, is the length of the vector $\mathbf{u} - \mathbf{v}$, i.e.

$$\text{dist}(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$$

If $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$, then

$$\begin{aligned} \text{dist}(\mathbf{u}, \mathbf{v}) &= \|\mathbf{u} - \mathbf{v}\| = \sqrt{(\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})} \\ &= \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2} \end{aligned}$$

Definition 5. Two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ are **orthogonal** if $\mathbf{u} \cdot \mathbf{v} = 0$.

We note that the zero vector is orthogonal to every vector in $\mathbb{R}^n$ because $\mathbf{0}^T\mathbf{v} = 0$ for all $\mathbf{v}$.

Theorem 2. (Pythagorean Theorem) Two vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

Definition 6. If a vector $\mathbf{z}$ is orthogonal to every vector in a subspace W of $\mathbb{R}^n$, then $\mathbf{z}$ is said to be **orthogonal to** W. The set of all vectors $\mathbf{z}$ that are orthogonal to W is called the **orthogonal complement** of W and is denoted by $W^\perp$.

The following are two facts about $W^\perp$, with W a subspace of $\mathbb{R}^n$:

1. A vector $\mathbf{x}$ is in $W^\perp$ if and only if $\mathbf{x}$ is orthogonal to every vector in a set that spans W.
2. $W^\perp$ is a subspace of $\mathbb{R}^n$.

Theorem 3. Let A be an $m \times n$ matrix. The orthogonal complement of the row space of A is the null space of A, and the orthogonal complement of the column space of A is the null space of A^T:

$$(\text{Row } A)^\perp = \text{Nul } A \text{ and } (\text{Col } A)^\perp = \text{Nul } A^T$$

If $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors in either $\mathbb{R}^2$ or $\mathbb{R}^3$, the formula relating the inner product and the angle between $\mathbf{u}$ and $\mathbf{v}$ is $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta$.

1.6.2 Orthogonal Sets

Definition 7. A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ in $\mathbb{R}^n$ is said to be an **orthogonal set** if each pair of distinct vectors from the set is orthogonal, that is $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ whenever $i \neq j$.

Theorem 4. If $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is an orthogonal set of nonzero vectors in $\mathbb{R}^n$, then S is linearly independent and therefore a basis for the subspace spanned by S.

Definition 8. An **orthogonal basis** for a subspace W of $\mathbb{R}^n$ is a basis for W that is also an orthogonal set.

Theorem 5. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ be an orthogonal basis for a subspace W of $\mathbb{R}^n$. For each $\mathbf{y}$ in W, the weights in the linear combination

$$\mathbf{y} = c_1\mathbf{u}_1 + \cdots + c_p\mathbf{u}_p$$

are given by

$$c_j = \frac{\mathbf{y} \cdot \mathbf{u}_j}{\mathbf{u}_j \cdot \mathbf{u}_j}$$

Given an orthogonal vector $\mathbf{u}$ in $\mathbb{R}^n$, consider the problem of decomposing a vector $\mathbf{y}$ in $\mathbb{R}^n$ into the sum of two vectors, one a multiple of $\mathbf{u}$ and the other orthogonal to $\mathbf{u}$, i.e.

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z}$$

where $\hat{\mathbf{y}} = \alpha\mathbf{u}$ for some scalar α and $\mathbf{z}$ is some vector orthogonal to $\mathbf{u}$.

Given any scalar α, let $\mathbf{z} = \mathbf{y} - \alpha\mathbf{u}$ so that the above equation is satisfied. Then $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $\mathbf{u}$ if and only if

$$0 = (\mathbf{y} - \alpha\mathbf{u}) \cdot \mathbf{u} = \mathbf{y} \cdot \mathbf{u} - (\alpha\mathbf{u}) \cdot \mathbf{u} = \mathbf{y} \cdot \mathbf{u} - \alpha(\mathbf{u} \cdot \mathbf{u})$$

That is, the equation above is satisfied with $\mathbf{z}$ orthogonal to $\mathbf{u}$ if and only if $\alpha = \frac{\mathbf{y} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}}$ and $\hat{\mathbf{y}} = \frac{\mathbf{y} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}}\mathbf{u}$.

Definition 9. The vector $\hat{\mathbf{y}}$ is called the **orthogonal projection of y onto u** and the vector $\mathbf{z}$ is called the **component of y orthogonal to u**.

If c is any nonzero scalar and if $\mathbf{u}$ is replaced by $c\mathbf{u}$ in the definition of $\hat{\mathbf{y}}$, then the orthogonal projection of $\mathbf{y}$ onto $c\mathbf{u}$ is exactly the same as the orthogonal projection of $\mathbf{y}$ onto $\mathbf{u}$. Hence, this projection is determined by the subspace L spanned by $\mathbf{u}$.

Definition 10. Sometimes $\hat{\mathbf{y}}$ is denoted by $\text{proj}_L\mathbf{y}$ and is called the **orthogonal projection of** y **onto** L. That is,

$$\hat{\mathbf{y}} = \text{proj}_L\mathbf{y} = \frac{\mathbf{y} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}}\mathbf{u}$$

Definition 11. A set $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is an **orthonormal set** if it is an orthogonal set of unit vectors. If W is the subspace spanned by such a set, then $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is an **orthonormal basis** for W (since the set is automatically linearly independent).

The simplest example of an orthonormal set is the standard basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ for $\mathbb{R}^n$.

When the vectors in an orthogonal set of nonzero vectors are normalized to have unit length, the new vectors will still be orthogonal, and hence the new set will be an orthonormal set.

Theorem 6. An $m \times n$ matrix U has orthonormal columns if and only if $U^T U = I$.

Theorem 7. Let U be an $m \times n$ matrix with orthonormal columns, and let $\mathbf{x}$ and $\mathbf{y}$ be in $\mathbb{R}^n$. Then

1. $\|U\mathbf{x}\| = \|\mathbf{x}\|$
2. $(U\mathbf{x}) \cdot (U\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$
3. $(U\mathbf{x}) \cdot (U\mathbf{y}) = 0$ if and only if $\mathbf{x} \cdot \mathbf{y} = 0$

Definition 12. An **orthogonal matrix** is a square invertible matrix U such that $U^{-1} = U^T$.

Any square matrix with orthonormal columns is an orthogonal matrix. Such a matrix must have orthonormal rows as well.

1.6.3 Orthogonal Projections

Whenever a vector $\mathbf{y}$ is written as a linear combination of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ in $\mathbb{R}^n$, the terms in the sum for $\mathbf{y}$ can be grouped into two parts so that $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{z}_1 + \mathbf{z}_2$$

where $\mathbf{z}_1$ is a linear combination of some of the $\mathbf{u}_i$ and $\mathbf{z}_2$ is a linear combination of the rest of the $\mathbf{u}_i$. This idea is particularly useful when $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is an orthogonal basis.

Theorem 8. (Orthogonal Decomposition Theorem) Let W be a subspace of $\mathbb{R}^n$. Then each $\mathbf{z}$ in $\mathbb{R}^n$ can be written uniquely in the form

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z}$$

where $\hat{\mathbf{y}}$ is in W and $\mathbf{z}$ is in $W^\perp$. In fact, if $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is any orthogonal basis of W, then

$$\hat{\mathbf{y}} = \frac{\mathbf{y} \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\mathbf{u}_1 + \cdots + \frac{\mathbf{y} \cdot \mathbf{u}_p}{\mathbf{u}_p \cdot \mathbf{u}_p}\mathbf{u}_p \text{ and } \mathbf{z} = \mathbf{y} - \hat{\mathbf{y}}$$

Theorem 9. The vector $\hat{\mathbf{y}}$ is called the **orthogonal projection of y onto** W and is often written as $\text{proj}_W \mathbf{y}$.

If $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is an orthogonal basis for W and if $\mathbf{y}$ is in $W = \text{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$, then $\text{proj}_W \mathbf{y} = \mathbf{y}$.

Theorem 10. (Best Approximation Theorem) Let W be a subspace of $\mathbb{R}^n$, let $\mathbf{z}$ be any vector in $\mathbb{R}^n$, and let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto W. Then $\hat{\mathbf{y}}$ is the closest point in W to $\mathbf{y}$, in the sense that

$$\|\mathbf{y} - \hat{\mathbf{y}}\| < \|\mathbf{y} - \mathbf{v}\|$$

for all $\mathbf{v}$ in W distinct from $\hat{\mathbf{y}}$.

Definition 13. The vector $\hat{\mathbf{y}}$ is called the **best approximation** to $\mathbf{y}$ by elements of W and the distance from $\mathbf{y}$ to $\mathbf{v}$, given by $\|\mathbf{y} - \mathbf{v}\|$ can be regarded as the error of using $\mathbf{y}$ in place of $\mathbf{y}$. This error is minimized when $\mathbf{v} = \hat{\mathbf{y}}$.

Theorem 11. If $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is an orthonormal basis for a subspace W of $\mathbb{R}^n$, then

$$\text{proj}_W \mathbf{y} = (\mathbf{y} \cdot \mathbf{u}_1)\mathbf{u}_1 + \cdots + (\mathbf{y} \cdot \mathbf{u}_p)\mathbf{u}_p$$

If $U = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{bmatrix}$, then

$$\text{proj}_W \mathbf{y} = UU^T\mathbf{y}$$

for all $\mathbf{y}$ in $\mathbb{R}^n$

1.6.4 The Gram-Schmidt Process

The Gram-Schmidt process is a simple algorithm for producing an orthogonal or orthonormal basis for any nonzero subspace of $\mathbb{R}^n$.

Theorem 12. Given a basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ for a nonzero subspace W of $\mathbb{R}^n$, define

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{x}_1 \\
\mathbf{v}_2 &= \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\mathbf{x}_1 \\
\mathbf{v}_3 &= \mathbf{x}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\mathbf{x}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{x}_2}{\mathbf{x}_2 \cdot \mathbf{x}_2}\mathbf{x}_2 \\
&\vdots \\
\mathbf{v}_p &= \mathbf{x}_1 - \frac{\mathbf{x}_p \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\mathbf{x}_1 - \frac{\mathbf{x}_p \cdot \mathbf{x}_2}{\mathbf{x}_2 \cdot \mathbf{x}_2}\mathbf{x}_2 - \frac{\mathbf{x}_p \cdot \mathbf{x}_{p-1}}{\mathbf{x}_{p-1} \cdot \mathbf{x}_{p-1}}\mathbf{x}_{p-1}
\end{aligned}$$

Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ is an orthogonal basis for W. In addition

$$\text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} = \text{Span}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \text{ for } 1 \leq k \leq p$$

The following is an example of the Gram-Schmidt process for

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Clearly, $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is linearly independent and thus is a basis for a subspace W of $\mathbb{R}^4$. To find an orthogonal basis, we perform the following steps.

Let $\mathbf{v}_1 = \mathbf{x}_1$ and $W_1 = \text{Span}\{\mathbf{x}_1\} = \text{Span}\{\mathbf{v}_1\}$.

Let $\mathbf{v}_2$ be the vector produced by subtracting from $\mathbf{x}_2$ its projection onto subspace W_1.

$$\begin{aligned}\mathbf{v}_2 &= \mathbf{x}_2 - \text{proj}_{W_1}\mathbf{x}_2 \\ &= \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 \\ &= \begin{bmatrix}0\\1\\1\\1\end{bmatrix} - \frac{3}{4}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} = \begin{bmatrix}-3/4\\1/4\\1/4\\1/4\end{bmatrix}\end{aligned}$$

Since $\mathbf{v}_2$ has fractional entries, we can scale it by a factor of 4 and replace $\{\mathbf{v}_1, \mathbf{v}_2\}$ by the orthogonal basis

$$\mathbf{v}_1 = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}, \mathbf{v}_2' = \begin{bmatrix}-3\\1\\1\\1\end{bmatrix}$$

Let $\mathbf{v}_3$ be the vector produced by subtracting from $\mathbf{x}_3$ its projection onto the subspace W_2. We use the orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2'\}$ to compute this projection onto W_2.

$$\text{proj}_{W_2}\mathbf{x}_3 = \frac{\mathbf{x}_3 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 + \frac{\mathbf{x}_3 \cdot \mathbf{v}_2'}{\mathbf{v}_2' \cdot \mathbf{v}_2'}\mathbf{v}_2' = \frac{2}{4}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} + \frac{2}{12}\begin{bmatrix}-3\\1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\2/3\\2/3\\2/3\end{bmatrix}$$

Then $\mathbf{v}_3$ is the component of $\mathbf{x}_3$ orthogonal to W_2, namely

$$\mathbf{v}_3 = \mathbf{x}_3 - \text{proj}_{W_2}\mathbf{x}_3 = \begin{bmatrix}0\\0\\1\\1\end{bmatrix} - \begin{bmatrix}0\\2/3\\2/3\\2/3\end{bmatrix} = \begin{bmatrix}0\\-2/3\\1/3\\1/3\end{bmatrix}$$

We end up with orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2', \mathbf{v}_3\}$ as an orthogonal basis for W:

$$\mathbf{v}_1 = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}, \mathbf{v}_2' = \begin{bmatrix}0\\2/3\\2/3\\2/3\end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix}0\\-2/3\\1/3\\1/3\end{bmatrix}$$

An orthonormal basis is constructed easily from an orthogonal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ by simply normalizing all the $\mathbf{v}_k$.

The orthonormal basis for the previous example is obtained to be:

$$\mathbf{v}_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -3/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 0 \\ -2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$$

If an $m \times n$ matrix A has linearly independent columns $\mathbf{x}_1, \ldots, \mathbf{x}_n$, then applying the Gram-Schmidt process (with normalizations) to $\mathbf{x}_1, \ldots, \mathbf{x}_n$ amounts to factoring A.

Theorem 13. (QR Factorization) If A is an $m \times n$ matrix with linearly independent columns, then A can be factored as $A = QR$, where Q is an $m \times n$ matrix whose columns form an orthonormal basis for Col A and R is an $n \times n$ upper triangular invertible matrix with positive entries on its diagonal.

We can find a QR factorization as follows for

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

The columns of A are the vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ above and taking the orthonormal basis for Col $A = \text{Span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ as the columns of Q, we have

$$Q = \begin{bmatrix} 1/2 & -3/\sqrt{12} & 0 \\ 1/2 & 1/\sqrt{12} & -2/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \end{bmatrix}$$

To find R, we observe that $Q^TQ = I$ and hence $Q^TA = Q^T(QR) = IR = R$, and so

$$\begin{aligned} R &= \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -3/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} \\ 0 & -2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 3/2 & 1 \\ 0 & 3/\sqrt{12} & 2/\sqrt{12} \\ 0 & 0 & 2/\sqrt{6} \end{bmatrix} \end{aligned}$$

1.6.5 Least Squares Problems

Definition 14. The general **least-squares problem** is to find an $\mathbf{x}$ that makes $\|\mathbf{b} - A\mathbf{x}\|$ as small as possible.

Definition 15. If A is $m \times n$ and $\mathbf{b}$ is in $\mathbb{R}^m$, a **least-squares solution** of $A\mathbf{x} = \mathbf{b}$ is an $\hat{\mathbf{x}}$ in $\mathbb{R}^n$ such that

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$$

for all $\mathbf{x}$ in $\mathbb{R}^n$.

No matter what $\mathbf{x}$ we select, the vector $A\mathbf{x}$ will necessarily be in the column space, Col A, so we seek an $\mathbf{x}$ that makes $A\mathbf{x}$ the closest point in Col A to $\mathbf{b}$.

Given A and $\mathbf{b}$ as above, we apply the Best Approximation Theorem to the subspace Col A. Let

$$\hat{\mathbf{b}} = \text{proj}_{\text{Col}A}\mathbf{b}$$

Because $\hat{\mathbf{b}}$ is in the column space of A, the equation $A\mathbf{x} = \hat{\mathbf{b}}$ is consistent and there is an $\hat{\mathbf{x}}$ in $\mathbb{R}^n$ such that

$$A\hat{\mathbf{x}} = \hat{\mathbf{b}}$$

Since $\hat{\mathbf{b}}$ is the closest point in Col A to $\mathbf{b}$, a vector $\hat{\mathbf{x}}$ is a least-squares solution of $A\mathbf{x} = \mathbf{b}$ if and only if $\hat{\mathbf{x}}$ satisfies $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$. Such an $\mathbf{x}$ in $\mathbb{R}^n$ is a list of weights that will build $\hat{\mathbf{b}}$ out of the columns of A.

Suppose $\hat{\mathbf{x}}$ satisfies $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$. By the Orthogonal Decomposition Theorem, the projection $\hat{\mathbf{b}}$ has the property that $\mathbf{b} - \hat{\mathbf{b}}$ is orthogonal to Col A, so $\mathbf{b} - A\hat{\mathbf{x}}$ is orthogonal to each column of A. If $\mathbf{a}_j$ is any column of A, then $\mathbf{a}_j \cdot (\mathbf{b} - A\hat{\mathbf{x}}) = 0$ and $\mathbf{a}_j^T(\mathbf{b} - A\hat{\mathbf{x}}) = 0$. Since each $\mathbf{a}_j^T$ is a row of A^T,

$$\begin{aligned} A^T(\mathbf{b} - A\hat{\mathbf{x}}) &= \mathbf{0} \\ A^T\mathbf{b} - A^TA\hat{\mathbf{x}} &= \mathbf{0} \\ A^TA\hat{\mathbf{x}} &= A^T\mathbf{b} \end{aligned}$$

These calculations show that the least-squares solution of $A\mathbf{x} = \mathbf{b}$ satisfies the equation

$$A^TA\mathbf{x} = A^T\mathbf{b}$$

The matrix equation above represents a system of equations called the **normal equations** for $A\mathbf{x} = \mathbf{b}$. A solution is often denoted by $\hat{\mathbf{x}}$.

Theorem 14. The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ coincides with the nonempty set of solutions of the normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$.

Theorem 15. Let A be an $m \times n$ matrix. The following statements are logically equivalent:

1. The equation $A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution for each $\mathbf{b}$ in $\mathbb{R}^m$.
2. The columns of A are linearly independent.
3. The matrix A^TA is invertible.

When these statements are true, the least-squares solution $\hat{\mathbf{x}}$ is given by

$$\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$$

Definition 16. When a least-squares solution $\hat{\mathbf{x}}$ is used to produce $A\hat{\mathbf{x}}$ as an approximation to $\mathbf{b}$, the distance from $\mathbf{b}$ to $A\hat{\mathbf{x}}$ is called the **least-squares error** of this approximation.

Theorem 16. Given an $m \times n$ matrix A with linearly independent columns, let $A = QR$ be a QR factorization of A. Then, for each $\mathbf{b}$ in $\mathbb{R}^m$, the equation $A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution, given by

$$\hat{\mathbf{x}} = R^{-1}Q^T\mathbf{b}$$

1.6.6 Inner Product Spaces

Definition 17. An **inner product** on a vector space V is a function that, to each pair of vectors $\mathbf{u}$ and $\mathbf{v}$ in V, associates a real number $\langle \mathbf{u}, \mathbf{v} \rangle$ and satisfies the following axioms, for all $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in V and all scalars c:

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
2. $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$
3. $\langle c\mathbf{u}, \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$
4. $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ and $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$

A vector space with an inner product is called an **inner product space**.

The vector space $\mathbb{R}^n$ with the standard inner product is an inner product space.

Definition 18. Let V be an inner product space, with the inner product denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$. We define the **length**, or **norm**, of a vector $\mathbf{v}$ to be the scalar

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$$

Equivalently, $\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle$.

Definition 19. A **unit vector** is one whose length is 1. The **distance between u and v** is $\|\mathbf{u} - \mathbf{v}\|$.Vectors $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Theorem 17. (Cauchy-Schwarz Inequality) For all $\mathbf{u}, \mathbf{v}$ in V,

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|$$

If $\mathbf{u} = \mathbf{0}$, then both sides are zero, and hence the inequality is true. If $\mathbf{u} \neq \mathbf{0}$, let W be the subspace spanned by $\mathbf{u}$. We recall that $\|c\mathbf{u}\| = |c|\|\mathbf{u}\|$ for any scalar c. Thus

$$\|\text{proj}_W \mathbf{v}\| = \left\| \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \right\| = \frac{|\langle \mathbf{v}, \mathbf{u} \rangle|}{|\langle \mathbf{u}, \mathbf{u} \rangle|} \|\mathbf{u}\| = \frac{|\langle \mathbf{v}, \mathbf{u} \rangle|}{\|\mathbf{u}\|^2} \|\mathbf{u}\| = \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|}{\|\mathbf{u}\|}$$

Since $\|\text{proj}_W \mathbf{v}\| \leq \|\mathbf{v}\|$, we have $\frac{|\langle \mathbf{u}, \mathbf{v} \rangle|}{\|\mathbf{u}\|} \leq \|\mathbf{v}\|$.

Theorem 18. (Triangle Inequality) For all $\mathbf{u}, \mathbf{v}$ in V,

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$$

Proof. We can show this using the following:

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|^2 &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \\ &\leq \|\mathbf{u}\|^2 + 2|\langle \mathbf{u}, \mathbf{v} \rangle| + \|\mathbf{v}\|^2 \\ &\leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| + \|\mathbf{v}\|^2 \\ &= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2 \end{aligned}$$

□

1.7 Symmetric Matrices and Quadratic Forms

1.7.1 Diagonalization of Symmetric Matrices

Definition 1. A **symmetric** matrix is a matrix A such that $A^T = A$. Such a matrix is necessarily square. Its main diagonal entries are arbitrary, but its other entries occur in pairs, on opposite sides of the main diagonal.

Theorem 1. If A is symmetric, then any two eigenvectors from different eigenspaces are orthogonal.

Definition 2. An $n \times n$ matrix A is said to be **orthogonally diagonalizable** if there are an orthogonal matrix P (with $P^{-1} = P^T$) and a diagonal matrix D such that

$$A = PDP^T = PDP^{-1}$$

Such a diagonalization requires n linearly dependent and orthonormal eigenvectors. If A is orthogonally diagonalizable, then

$$A^T = (PDP^T)^T = P^{TT} D^T P^T = PDP^T = A$$

Definition 3. An $n \times n$ matrix A is orthogonally diagonalizable if and only if A is a symmetric matrix.

Theorem 2. (Spectral Theorem for Symmetric Matrices) An $n \times n$ symmetric matrix has the following properties:

1. A has n real eigenvalues, counting multiplicities.
2. The dimension of the eigenspace for each eigenvalue λ equals the multiplicity of λ as a root of the characteristic equation.
3. The eigenspaces are mutually orthogonal, in the sense that eigenvectors corresponding to different eigenvalues are orthogonal.
4. A is orthogonally diagonalizable.

Suppose $A = PDP^{-1}$, where the columns of P are orthonormal eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of A and the corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ are in the diagonal matrix D. Then, since $P^{-1} = P^T$, we have

$$\begin{aligned} A &= PDP^T \\ &= \begin{bmatrix} \mathbf{u}_1 & \cdots \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 \mathbf{u}_1 & \cdots & \lambda_n \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix} \\ &= \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \cdots + \lambda_n \mathbf{u}_n \mathbf{u}_n^T \end{aligned}$$

Definition 4. This representation of A is called a **spectral decomposition** of A because it breaks up A into pieces determined by the spectrum (eigenvalues) or A.

Each term is an $n \times n$ matrix of rank 1. For example, every column of $\lambda_1 \mathbf{u}_1 \mathbf{u}_1^T$ is a multiple of $\mathbf{u}_1$. Furthermore, each matrix $\mathbf{u}_j \mathbf{u}_j^T$ is a **projection matrix**, since for each $\mathbf{x}$ in $\mathbb{R}^n$, the vector $(\mathbf{u}_j \mathbf{u}_j^T)\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the subspace spanned by $\mathbf{u}_j$.

The following is a spectral decomposition of the matrix A with the orthogonal diagonalization

$$A = \begin{bmatrix} 7 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 2/\sqrt{5} & -1/\sqrt{5} \\ 1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 8 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix}$$

Denoting the columns of P by $\mathbf{u}_1$ and $\mathbf{u}_2$, we have $A = 8\mathbf{u}_1 \mathbf{u}_1^T + 3\mathbf{u}_2 \mathbf{u}_2^T$.

To verify this decomposition of A, we compute the following:

$$\begin{aligned} \mathbf{u}_1 \mathbf{u}_1^T &= \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 4/5 & 2/5 \\ 2/5 & 1/5 \end{bmatrix} \\ \mathbf{u}_2 \mathbf{u}_2^T &= \begin{bmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} -1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} = \begin{bmatrix} 1/5 & -2/5 \\ -2/5 & 4/5 \end{bmatrix} \\ 8\mathbf{u}_1 \mathbf{u}_1^T + 3\mathbf{u}_2 \mathbf{u}_2^T &= \begin{bmatrix} 32/5 & 16/5 \\ 16/5 & 8/5 \end{bmatrix} + \begin{bmatrix} 3/5 & -6/5 \\ -6/5 & 12/5 \end{bmatrix} = \begin{bmatrix} 7 & 2 \\ 2 & 4 \end{bmatrix} = A \end{aligned}$$

1.7.2 Quadratic Forms

Definition 5. A **quadratic form** on $\mathbb{R}^n$ is a function Q defined on $\mathbb{R}^n$ whose value at a vector $\mathbf{x}$ in $\mathbb{R}^n$ can be computed by an expression of the form $Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, where A is an $n \times n$ symmetric matrix. The matrix A is called the **matrix of the quadratic form**.

The simplest example of a nonzero quadratic from is $Q(\mathbf{x}) = \mathbf{x}^T I\mathbf{x} = \|\mathbf{x}\|^2$.

Definition 6. If $\mathbf{x}$ represents a variable vector in $\mathbb{R}^n$, then a **change of variable** is an equation of the form

$$\mathbf{x} = P\mathbf{y} \iff \mathbf{y} = P^{-1}\mathbf{x}$$

where P is an invertible matrix and $\mathbf{y}$ is a new variable vector in $\mathbb{R}^n$. Here $\mathbf{y}$ is the coordinate vector of $\mathbf{x}$ relative to the basis of $\mathbb{R}^n$ determined by the columns of P.

If the change of variable is made in a quadratic form $\mathbf{x}^T A\mathbf{x}$, then

$$\mathbf{x}^T A\mathbf{x} = (P\mathbf{y})^T A(P\mathbf{y}) = \mathbf{y}^T P^T AP\mathbf{y} = \mathbf{y}^T (P^T AP)\mathbf{y}$$

and the new matrix of the quadratic form is $P^T AP$.

Since A is symmetric, we are guaranteed an orthogonal matrix P such that $P^T AP$ is a diagonal matrix D and the quadratic form becomes $\mathbf{y}^T D\mathbf{y}$.

Theorem 3. (Principle Axes Theorem) Let A be an $n \times n$ symmetric matrix. Then there is an orthogonal change of variable, $\mathbf{x} = P\mathbf{y}$, that transforms the quadratic form $\mathbf{x}^T A\mathbf{x}$ into a quadratic form $\mathbf{y}^T D\mathbf{y}$ with no cross-product term.

Definition 7. The columns of P in the theorem are called the **principal axes** of the quadratic form $\mathbf{x}^T A\mathbf{x}$. The vector $\mathbf{y}$ is the coordinate vector of $\mathbf{x}$ relative to the orthonormal basis of $\mathbb{R}^n$ given by these principal axes.

Definition 8. A quadratic form Q is

1. **positive definite** if $Q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$,
2. **positive semidefinite** if $Q(\mathbf{x}) \geq 0$ for all $\mathbf{x} \neq 0$,
3. **negative definite** if $Q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$,
4. **negative semidefinite** if $Q(\mathbf{x}) \leq 0$ for all $\mathbf{x} \neq 0$,
5. **indefinite** if $Q(\mathbf{x})$ assumes both positive and negative values.

Theorem 4. (Quadratic Forms and Eigenvalues) Let A be an $n \times n$ symmetric matrix. Then a quadratic form $\mathbf{x}^T A\mathbf{x}$ is

1. positive definite if and only if the eigenvalues of A are all positive,
2. negative definite if and only if the eigenvalues of A are all negative, or
3. indefinite if and only if A has both positive and negative eigenvalues.

1.7.3 Singular Value Decomposition

A factorization $A = QDP^{-1}$ is possible for any $m \times n$ matrix A. The singular value decomposition is a special factorization of this type. It is based on the following property of the ordinary diagonalization that can be imitated for rectangular matrices: the absolute values of the eigenvalues of a symmetric matrix A measure the amounts that A stretches or shrinks certain vectors (the eigenvectors). If $A\mathbf{x} = \lambda\mathbf{x}$ and $\|\mathbf{x}\| = 1$, then

$$\|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\| = |\lambda|$$

Let A be an $m \times n$ matrix. Then A^TA is symmetric and can be orthogonally diagonalized. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of A^TA, and let $\lambda_1, \ldots, \lambda_n$ be the associated eigenvalues of A^TA. Then, for $1 \leq i \leq n$,

$$\begin{aligned}\|A\mathbf{v}_i\|^2 &= (A\mathbf{v}_i)^TA\mathbf{v}_i = \mathbf{v}_i^TA^TA\mathbf{v}_i \\ &= \mathbf{v}_i^T(\lambda_i\mathbf{v}_i) \\ &= \lambda_i\end{aligned}$$

So the eigenvalues of A^TA are all nonnegative. By renumbering, if necessary, we may assume that the eigenvalues are arranged so that

$$\lambda_1 \geq \cdots \geq \lambda_n \geq 0$$

Definition 9. The **singular values** of A are the square roots of the eigenvalues of A^TA, denoted by $\sigma_1, \ldots, \sigma_n$, and they are arranged in decreasing order. That is, $\sigma_i = \sqrt{\lambda_1}$ for $1 \leq i \leq n$. The singular values of A are the lengths of the vectors $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$.

Theorem 5. Suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A^TA, arranged so that the corresponding eigenvalues of A^TA satisfy $\lambda_1 \geq \cdots \geq \lambda_n$, and suppose that A has r nonzero singular values. Then $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is an orthogonal basis for Col A, and rank $A = r$.

The decomposition of A involves an $m \times n$ diagonal matrix Σ of the form

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

where D is an $r \times r$ diagonal matrix for some r not exceeding the smaller of m and n.

Theorem 6. (Singular Value Decomposition) Let A be an $m \times n$ matrix with rank r. Then there exists an $m \times n$ matrix Σ for which the diagonal entries in D are the first r singular values of A, $\sigma_1 \geq \cdots \geq \sigma_r > 0$, and there exist an $m \times m$ orthogonal matrix U and an $n \times n$ orthogonal matrix V such that

$$A = U\Sigma V^T$$

Definition 10. Any factorization $A = U\Sigma V^T$, with U and V orthogonal, Σ as above, and positive diagonal entries in D, is called a **singular value decomposition (SVD)** of A. The matrices U and V are not uniquely determined by A, but the diagonal entries of Σ are necessarily the singular values of A. The columns of U in such a decomposition are called the **left singular vectors** of A, and the columns of V are called **right singular vectors** of A.

The four fundamental subspaces and the concept of singular values provide the final statements of the Invertible Matrix Theorem.

Theorem 7. (Invertible Matrix Theorem) Let A be an $n \times n$ matrix. Then the following statements are each equivalent:

1. A is an invertible matrix.
2. A is row equivalent to the $n \times n$ identity matrix.
3. A has n pivot positions.
4. The equation $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
5. The columns of A form a linearly independent set.
6. The linear transformation $x \mapsto A\mathbf{x}$ is one-to-one.
7. The equation $A\mathbf{x} = \mathbf{b}$ has at least one solution for each $\mathbf{b}$ in $\mathbb{R}^n$.
8. The columns of A span $\mathbb{R}^n$.
9. The linear transformation $x \mapsto A\mathbf{x}$ maps $\mathbb{R}^n$ onto $\mathbb{R}^n$.
10. There is an $n \times n$ matrix C such that $CA = I$.
11. There is an $n \times n$ matrix D such that $AD = I$.
12. A^T is an invertible matrix.
13. The columns of A form a basis of $\mathbb{R}^n$.
14. Col $A = \mathbb{R}^n$
15. dim Col $A = n$
16. rank $A = n$
17. Nul $A = \{\mathbf{0}\}$
18. dim Nul $A = 0$
19. The number 0 is not an eigenvalue of A.
20. The determinant of A is not zero.
21. $(\text{Col } A)^\perp = \{\mathbf{0}\}$
22. $(\text{Nul } A)^\perp = \mathbb{R}^n$
23. Row $A = \mathbb{R}^n$
24. A has n nonzero singular values.

Chapter 2

Probability

2.1 Combinatorics

2.1.1 Basic Rules of Counting

We begin by building up ways that allow us to count a wide range of scenarios.

Suppose you have to make a two-stage decision: first, you must make one choice out of m total choices (where $m > 0$); then, regardless of your first choice, you must make one choice out of n total choices (where $n > 0$).

Definition 1. The **multiplication rule** states that when there are m choices at the first stage and n choices at the second stage, then there are a total of mn possibilities.

The multiplication rule extends easily to the case when there are more than two stages. If there are k stages and at each stage, there are n_i possible choices (for $i \in \{1, \ldots, k\}$), then the total number of possibilities is $\prod_{i=1}^{k} n_i$.

Definition 2. The **bijection principle** states that if there is a bijection between two sets, then the two sets have the same number of elements.

We introduce an analogy that we will use repeatedly with balls and bins. Let m, n be positive integers. We throw n balls into m bins and would like to count the number of ways this can happen. First, we assume that all of the balls and bins are distinguishable, i.e. the balls are numbered. We can thus think of the process of throwing the n balls as an n-stage decision process, where at each stage, each ball has m choices for its destination bin. This allows us to use the multiplication rule.

The number of ways to throw n distinguishable balls into m distinguishable bins is m^n.

Definition 3. Let n be a positive integer. A **permutation** of $\{1, \ldots, n\}$ is a way of rearranging the numbers $(1, \ldots, n)$, i.e. a permutation σ is a bijection $\sigma : \{1, \ldots, n\} \leftrightarrow \{1, \ldots, n\}$.

To count the number of permutations, we can again think of building a permutation as an n-stage decision process, where at the i^{th} step (for $i \in \{1, \ldots, n\}$), we are deciding on the value of $\sigma(i)$. Starting at $\sigma(1)$, we can choose anything in $\{1, \ldots, n\}$, so there are n choices. For $\sigma(2)$, we can choose anything in $\{1, \ldots, n\}$ except $\sigma(1)$, so there are $n-1$ choices. Similarly, for $\sigma(i)$, we have $n - (i-1)$ choices. Putting it together, the total number of possibilities is $\prod_{i=1}^{n}(n-(i-1)) = \prod_{i=1}^{n} i = 1 \cdot 2 \cdots (n-1) \cdot n$.

Definition 4. We define $n! = \prod_{i=1}^{n} i$ to be n **factorial** for positive integers n where $0! = 1$.

The number of permutations of $\{1, \ldots, n\}$ is

$$n! = \prod_{i=1}^{n} i = n \cdot (n-1) \cdots 2 \cdot 1$$

Another way to think about permutations is that we are counting the number of ordered subsets of size n from $\{1, \ldots, n\}$. Extending this, we can try to count the number of ordered subsets of size k from $\{1, \ldots, n\}$ where $k \in \{1, \ldots, n\}$. We can follow the same procedure as above, except instead of having n stages, we have k stages.

The number of ordered subsets of size k from $\{1, \ldots n\}$ is

$$\prod_{i=1}^{k}(n - (i-1)) = \frac{n!}{(n-k)!}$$

Next, we count the number of unordered subsets of size k from $\{1, \ldots, n\}$. We already know the number of ordered subsets of size k is $n!/(n-k)!$. We observe that each unordered subset gives rise to exactly $k!$ ordered subsets. Therefore, the number of unordered subsets must be $k!$ times less than the number of ordered subsets, giving us $n!/(k!(n-k)!)$.

Definition 5. We denote $\binom{n}{k}$ to be the binomial coefficient. Notably, $\binom{n}{k} = \binom{n}{n-k}$.

The number of unordered subsets of size k from $\{1, \ldots, n\}$ is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

We return to the analogy of throwing balls into bins and now assume that the balls are indistinguishable. To count the number of configurations, we use a technique called **stars and bars**. The idea is to construct a bijection between the set of all configurations of n indistinguishable balls in m distinguishable bins with the set of strings of length $n + m - 1$ from the alphabet $\{*, |\}$. We represent the n balls using n star symbols and the $m-1$ spaces between the bins using $m - 1$ bar symbols.

The following is the bijection for $m = n = 3$:

balls and bins configuration	stars and bars
3 balls in bin 1	***||
2 balls in bin 1, 1 ball in bin 2	**|*|
2 balls in bin 1, 1 ball in bin 3	**||*
1 ball in bin 1, 2 balls in bin 2	*|**|
1 ball in bin 1, 1 ball in bin 2, 1 ball in bin 3	*|*|*
1 ball in bin 1, 2 balls in bin 3	*||**
3 balls in bin 2	|***|
2 balls in bin 2, 1 ball in bin 3	|**|*
1 ball in bin 2, 2 balls in bin 3	|*|**
3 balls in bin 3	||***

Therefore, we only need to count the number of strings of length $n+m-1$ with exactly n stars and $m-1$ bars. Using what we found earlier, we find this to be $\binom{n+m-1}{n} = \binom{n+m-1}{m-1}$.

The number of ways to throw n indistinguishable balls into m distinguishable bins is

$$\binom{n+m-1}{n} = \binom{n+m-1}{m-1}$$

2.1.2 Combinatorial Proofs

We have seen that the number of bit strings of length n is 2^n. We have also seen that the number of bit strings of length n with exactly m ones is $\binom{n}{m}$. Therefore, if we sum over all possible values of m, then we will have the total number of bit strings of length n, i.e. $\sum_{m=0}^{n} \binom{n}{m} = 2^n$. This is known as a **combinatorial proof**.

Theorem 1. (Binomial Theorem) Let n be a positive integer. Then

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Definition 6. (Pascal's Identity) For positive integers $k \leq n$, $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.

Proof. Let X be a set with n elements and let $x \in X$. To choose a subset of size k from X, there are $\binom{n-1}{k-1}$ ways to choose a subset including x and $\binom{n-1}{k}$ ways to choose a subset not including x. Therefore, we have that $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.

We can also show the equivalence algebraically:

$$\begin{aligned}
\binom{n-1}{k-1} + \binom{n-1}{k} &= \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-k-1)!} \\
&= (n-1)!\left(\frac{k}{k!(n-k)!} + \frac{n-k}{k!(n-k)!}\right) \\
&= (n-1)! \cdot \frac{n}{k!(n-k)!} \\
&= \frac{n!}{k!(n-k)!} \\
&= \binom{n}{k}
\end{aligned}$$

□

2.1.3 Inclusion-Exclusion Principle

Let p and q be distinct prime numbers. To count the number of elements in the set $\{1, \ldots, pq\}$ that are relatively prime to both p and q, we can first find the numbers which are not relatively prime to p, i.e. the multiples of p, and likewise the numbers which are not relatively prime to q, i.e. the multiples of q. However, we must also find the numbers which are not relatively prime to both p and q, i.e. the multiples of pq, which would have been double counted. Hence, the set contains $p + q - 1$ numbers which are not relatively prime to p or q. Therefore, the set contains $pq - (p + q - 1)$ numbers which are relatively prime to p and q. This illustrates the principle that $|A \cup B| = |A| + |B| - |A \cap B|$. The following is the general principle.

Theorem 2. The **inclusion-exclusion principle** states that for finite sets $A_1, \ldots, A_n$,

$$\begin{aligned}\left|\bigcup_{i=1}^{n} A_i\right| &= \sum_{i=1}^{n}|A_i| - \sum_{i<j}|A_i \cap A_j| + \sum_{i<j<k}|A_i \cap A_j \cap A_k| - \cdots + (-1)^{n-1}|A_i \cap \cdots \cap A_n| \\ &= \sum_{k=1}^{n}(-1)^{k+1} \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \left(\bigcap_{i \in I} A_i\right)\end{aligned}$$

2.2 Probability Theory

2.2.1 Probability Axioms

A wide variety of phenomena in the world can be modeled by the following mathematical objects:

1. Ω, the **probability space**, a set of all possible outcomes of interest
2. subsets of the probability space, called **events** (sometimes denoted $\mathcal{F}$)
3. a function that assigns values to sets, called a **probability measure** $\mathbb{P}$

Some properties that our probability measure should satisfy are:

1. $\mathbb{P}(\varnothing) = 0$, i.e. the likelihood of nothing happening is zero, and
2. $\mathbb{P}(\Omega) = 1$, in order to restrict probability values to the range $[0, 1]$, and
3. $\mathbb{P}(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i)$, i.e. **countable additivity**, which states that if for each $i \in \mathbb{N}$ we have an event A_i, such that the events A_i are pairwise disjoint, then the probabilities of the events must add.

Definition 1. If A_1 and A_2 are disjoint events, then $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$. This is known as **finite additivity** and follows the above.

2.2.2 Fundamental Probability Facts

From the axioms, we can derive other important rules.

Definition 2. For any event A, we denote A^{c} to be the **complement** of A.

We begin by writing $\Omega = A \cup A^{\mathsf{c}}$ and obtain the following by applying finite additivity:

$$\begin{aligned} 1 &= \mathbb{P}(\Omega) \\ 1 &= \mathbb{P}(A \cup A^{\mathsf{c}}) \\ 1 &= \mathbb{P}(A) + \mathbb{P}(A^{\mathsf{c}}) \\ \mathbb{P}(A^{\mathsf{c}}) &= 1 - \mathbb{P}(A) \end{aligned}$$

Definition 3. The next is known as **subadditivity**, which states that if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$. We can show this by observing that if $A \subseteq B$, then $B = A + (B \setminus A)$, where A and $B \setminus A$ are disjoint sets, giving us $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \Rightarrow \mathbb{P}(B) \geq \mathbb{P}(A)$.

Definition 4. The next is the **inclusion-exclusion principle**, which states that for two events A, B,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

and in general,

$$\begin{aligned} \mathbb{P}(\bigcup_{i=1}^{n} A_i) &= \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} \mathbb{P}(A_i \cap \cdots \cap A_n) \\ &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} \mathbb{P}\Big(\bigcap_{i \in I} A_i\Big) \end{aligned}$$

If we take the inclusion-exclusion principle for two events and drop the last term on the right, then we have the simple equality $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. We can extend this as follows.

Definition 5. The **union bound** states that if n is a positive integer and $A_1, \ldots, A_n$ are events, then

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$$

Proof. We can show this using induction. The base case when $n = 1$ is trivial. For $n > 1$, suppose that the result holds for n disjoint events and let $A_1, \ldots, A_{n+1}$ be disjoint. By applying finite additivity to the events $A_1 \cup \cdots \cup A_n$ and A_{n+1} (which are disjoint), we have $\mathbb{P}(A_1 \cup \cdots \cup A_{n+1}) = \mathbb{P}(A_1 \cup \cdots \cup A_n) + \mathbb{P}(A_{n+1})$. Applying the inductive claim to the first term, we conclude that $\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$. □

In fact, we can extend the union bound to an infinite number of events. Let $A_1, A_2, A_3, \ldots$ be a countably infinite sequence of events,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Proof. We can define $A'_1 = A_1$ and for each $i = 2, 3, \ldots$, define $A'_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j$. Therefore, for each positive integer i, we have $A'_i \subseteq A_i$ and so $\mathbb{P}(A'_i) \leq \mathbb{P}(A_i)$. We also notice that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} A'_i$ and the sequence of events $A'_1, A'_2, A'_3, \ldots$ is disjoint, so we can apply the countable additivity axiom, concluding that

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \mathbb{P}(\bigcup_{i=1}^{\infty} A'_i) = \sum_{i=1}^{\infty} \mathbb{P}(A'_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

□

2.2.3 Discrete Probability

We now consider the case where Ω is countable. We recall that the probability measure $\mathbb{P}$ is a function which assigns real numbers to sets of outcomes. In the case of discrete probability, the probability of any event is completely determined once we specify the probability of each outcome, i.e.

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$$

Definition 6. We say that Ω is a **uniform probability space** when every outcome is equally likely to occur, i.e. $\mathbb{P}(\omega) = 1/|\Omega|$.

The advantage of having a uniform probability space is that we may employ methods of combinatorics to compute probabilities, $\mathbb{P}(A) = |A|/|\Omega|$, where to compute the probability of the event A, we simply count the number of ways in which A is achieved and divide by the total number of elements in Ω.

2.2.4 Conditional Probability

Definition 7. The **law of total probability** states the following. Let n be a positive integer and suppose that the events $A_1, \ldots, A_n$ partition the sample space, that is, $A_1 \cup \cdots \cup A_n = \Omega$. Then, to find the probability of an event B, we can write

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i)$$

Definition 8. The idea behind **conditional probability** is that events can affect one another. We write $\mathbb{P}(B \mid A)$ to denote the probability of B given A.

The new law is a true probability law, i.e. it must satisfy the three probability axioms, and we would like $\mathbb{P}(A \mid A) = 1$. Let A be an event with $\mathbb{P}(A) > 0$. Then

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Provided that $\mathbb{P}(B) > 0$ as well, the expression above can also be written as

$$\mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$$

Definition 9. The equation $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A)$ is often known as the **chain rule** or **product rule** and states that we can directly apply our cause-and-effect knowledge by multiplying the unconditional probability $\mathbb{P}(A)$ with the conditional probability $\mathbb{P}(B \mid A)$ to obtain the probability that both events A and B occur.

We use the law of total probability and the definition of conditional probability to obtain the following.

Definition 10. Bayes' Rule is stated as the following equation:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A \cap B) + \mathbb{P}(A^{\mathsf{c}} \cap B)} = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^{\mathsf{c}})\mathbb{P}(A^{\mathsf{c}})}$$

More generally, if n is a positive integer and $A_1, \ldots, A_n$ partition the sample space, then

$$\mathbb{P}(A_k \mid B) = \frac{\mathbb{P}(B \mid A_k)\mathbb{P}(A_k)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}$$

for $k = 1, \ldots, n$.

Given multiple possible explanations for an observation we have made, we can discern the most likely explanation. This is called **probabilistic inference**.

We can also interpret conditional probability in the framework of taking an old probability law and producing a new probability law after observing an event, in the sense of an **update rule**.

2.2.5 Independence

Definition 11. If we observe an event A, but knowing that A occurs tells us exactly nothing about whether B will occur, then we say that A and B are **independent**, i.e.

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Another way of viewing independence, when $\mathbb{P}(A) > 0$, is $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

If we view the independence of A and B as the information about B that one receives from observing A (or vice versa), then we should have that A^{c} and B are also independent, A and B^{c} are also independent, etc.

Definition 12. We say that $A_1, \ldots, A_n$ are **pairwise independent** if every pair of events is independent.

In the case where n is a positive integer, we can have an even stronger statement.

Definition 13. We say that $A_1, \ldots, A_n$ are **mutually independent** if we have two sets of events, such that the sets are disjoint, and any combination of events from the first set is independent of any combination of events from the second event, i.e.

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i'\Big) = \prod_{i=1}^{n} \mathbb{P}(A_i')$$

where each A_i' is allowed to be either A_i or Ω.

For an infinite set of events $\{A_\alpha\}_{\alpha \in \mathcal{A}}$ for some indexing set $\mathcal{A}$, we say $\{\mathcal{A}_\alpha\}_{\alpha \in \mathcal{A}}$ is independent if every finite subcollection of the events is independent.

Definition 14. We introduce the notion of correlation:

1. We say that events A and B are **positively correlated** if $\mathbb{P}(A \cap B) > \mathbb{P}(A)\mathbb{P}(B)$.
2. We say that events A and B are **negatively correlated** if $\mathbb{P}(A \cap B) < \mathbb{P}(A)\mathbb{P}(B)$.

If $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then A are B are independent.

Intuitively, if $\mathbb{P}(A \cap B)$ is large relative to $\mathbb{P}(A)\mathbb{P}(B)$, then events A and B tend to occur together. Observing one of A or B increases the likelihood of observing the other. If $\mathbb{P}(A \cap B)$ is small relative to $\mathbb{P}(A)\mathbb{P}(B)$, then observing one of A or B decreases the likelihood of observing the other. An extreme example is when A and B are mutually exclusive, i.e. disjoint.

2.3 Discrete Random Variables and Inequalities

2.3.1 Random Variables

Definition 1. A **random variable** is a function $X : \Omega \to \mathbb{R}$ that assigns a real number to every outcome ω in the probability space. We typically denote random variables by capital letters and view random variables as numerical outcomes associated with a random experiment.

We define the sum of random variables $X + Y$ to be the random variable that maps ω to $X(\omega) + Y(\omega)$. Similarly, we define the product of random variables XY to be the random variable that maps ω to $X(\omega)Y(\omega)$.

More generally, let n be any positive integer and let $f : \mathbb{R}^n \to \mathbb{R}$ be any function. Then $f(X_1, \ldots, X_n)$ is defined to be the random variable that maps ω to $f(X_1(\omega), \ldots, X_n(\omega))$.

Random variables assign probabilities to real numbers. We can see this while considering the example of tossing two fair coins. The sample space of this experiment is $\Omega = \{HH, HT, TH, TT\}$ and we can define the random variable X to be the number of heads we see in the two coin tosses, i.e.

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0$$

Therefore, the probability that X takes on a particular value is

$$\mathbb{P}(X = 0) = \mathbb{P}(\{TT\}) = \frac{1}{4}$$
$$\mathbb{P}(X = 1) = \mathbb{P}(\{HT, TH\}) = \frac{1}{2}$$
$$\mathbb{P}(X = 2) = \mathbb{P}(\{HH\}) = \frac{1}{4}$$

We can thus say that X assigns the probability $1/4$ to the real number 0, the probability $1/2$ to the real number 1, and the probability $1/4$ to the real number 2.

Definition 2. In this way, X induces a probability measure on the real line. We call this the **distribution** of X. The distribution of X satisfies the probability axioms.

An advantage of random variables is that we can often forget about the underlying probability space Ω and focus our attention on the distribution of X. For a discrete random variable X, we can specify the distribution of X by simply giving the probabilities $\mathbb{P}(X = x)$ for all x in the range of X without reference to the original probability space Ω.

Definition 3. Equivalently, we can describe a probability distribution by its **cumulative distribution function** or CDF, which is defined as $F_X(x) = \mathbb{P}(X \leq x)$. The CDF contains exactly the same information as the distribution of X. To observe this, we can recover the **probability distribution function** or PDF using the following formula:

$$\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X \leq x - 1)$$

for all integers x, assuming X takes on integer values.

Definition 4. The **joint distribution** of two random variables X and Y is the probability distribution $\mathbb{P}(X = x, Y = y)$ for all possible pairs of values (x, y). The joint distribution must satisfy the normalization condition, i.e.

$$\sum_x \sum_y \mathbb{P}(X = x, Y = y) = 1$$

Definition 5. We can recover the distribution of X (resp. Y) separately, or the **marginal distribution** of X (resp. Y), by summing over all possible values of Y (resp. X), i.e.

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y)$$

$$\mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y)$$

The joint distribution contains all of the information about both X and Y. From the joint distribution, we can recover the marginal distributions of X and Y. The converse is not true; the marginal distributions are usually not sufficient to recover the joint distribution, since the joint distribution contains information about the dependence of X and Y.

Definition 6. We say that two discrete random variables are independent if $\forall x, y \in \mathbb{R}$,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

If X and Y are independent, then we can write their joint probability as a product of their marginal probabilities.

2.3.2 Expectation

Definition 7. The **expectation** or **expected value** of a discrete random variable X is defined to be

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) = \sum_x x\mathbb{P}(X = x)$$

The expected value is interpreted as the long-run average of an experiment in which one measures the values of X, i.e. if $X_1, X_2, X_3, \ldots$ are independent copies of X, then

$$\lim_{N \to \infty} \frac{X_1 + \cdots + X_N}{N} \to \mathbb{E}[X]$$

Theorem 1. (Linearity of Expectation) Suppose X, Y are random variables, $a \in \mathbb{R}$ is a constant, and c is the constant random variable. Then

1. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
2. $\mathbb{E}[aX + c] = a\mathbb{E}[X] + c$

If X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is a function, then $f(X)$ is a random variable and its expectation is

$$\mathbb{E}[f(X)] = \sum_{\omega \in \Omega} f(X(\omega))\mathbb{P}(\omega) = \sum_x f(x)\mathbb{P}(X = x)$$

This can be extended to functions of multiple random variables as follows:

$$\mathbb{E}[f(X_1, \ldots, X_n)] = \sum_{x_1, \ldots x_n} f(x_1, \ldots, x_n)\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

Theorem 2. (Expectation of Independent Random Variables) Let X and Y be independent random variables. Then the random variable XY satisfies

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

Proof. We can manipulate the formula for expectation as follows:

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x,y} xy\mathbb{P}(X = x, Y = y) \\
&= \sum_{x,y} xy\mathbb{P}(X = x)\mathbb{P}(Y = y) \\
&= \Big(\sum_{x} x\mathbb{P}(X = x)\Big)\Big(\sum_{y} y\mathbb{P}(Y = y)\Big) \\
&= \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

□

Theorem 3. (Tail Sum Theorem) Let X be a random variable on values in $\mathbb{N}$. Then

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x)$$

Proof. We can manipulate the formula for expectation as follows:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \sum_{k=1}^{\infty}\sum_{x=1}^{k} \mathbb{P}(X = k) \\
&= \sum_{k=1}^{\infty}\sum_{k=x}^{\infty} \mathbb{P}(X = k) = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x)
\end{aligned}$$

□

This formula is known as the **tail sum formula** because we compute the expectation by summing over the tail probabilities of the distribution.

2.3.3 Variance

The variance is a measure of the spread of a distribution and how unpredictable your results are. One possible quantity we can study is $X - \mathbb{E}[X]$, but we notice that the expectation $\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$ no matter the distribution, so studying the average of the differences is not interesting. Another possible quantity we can study is $|X - \mathbb{E}[X]|$, but this makes it difficult to solve problems analytically.

Definition 8. The **variance** of a probability distribution is

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Definition 9. We often denote the mean of the probability distribution as μ and the variance as σ^2. We call $\sigma = \sqrt{\text{var}[X]}$ the **standard deviation** of X.

The standard deviation is useful because it has the same units as X and thus allows for easier comparison.

We can use linearity of expression to compute a formula for variance:

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

Thus, we find the explicit formula for variance to be

$$\text{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Theorem 4. Let X be a random variable, $a \in \mathbb{R}$ be a constant, and c be the constant random variable. Then

$$\text{var}[aX + c] = a^2\text{var}[X]$$

Taking the square root, the following statement holds for standard deviations:

$$\sigma_{aX+c} = |a|\sigma_X$$

Theorem 5. Let X and Y be random variables. Then

$$\text{var}[X + Y] = \text{var}[X] + \text{var}[Y] + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])$$

It follows that when X and Y are independent,

$$\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$$

2.3.4 Discrete Probability Distributions

We now discuss the common discrete probability distributions.

Definition 10. Consider the **uniform distribution** over $\{1, \ldots, n\}$, denoted as Uniform$\{1, \ldots n\}$. Since each element of the set is equally likely to be chosen, the probability distribution is

$$\mathbb{P}(X = x) = \frac{1}{n}, \quad x \in \{1, \ldots, n\}$$

The expectation of the uniform distribution is

$$\mathbb{E}[x] = \sum_{x=1}^{n} x \cdot \frac{1}{n} = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

We can first compute

$$\mathbb{E}[X^2] = \sum_{x=1}^{n} x^2 \cdot \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

The variance of the uniform distribution is

$$\text{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

Definition 11. The **Bernoulli distribution with parameter** p, $p \in [0,1]$, denoted Bernoulli(p), is a distribution that describes the result of performing one experiment which succeeds with probability p. The probability distribution is

$$P(X = x) = \begin{cases} 1-p & x = 0, \\ p & x = 1, \\ 0 & \text{otherwise} \end{cases}$$

The expectation of the Bernoulli(p) distribution is

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = 0 \cdot (1-p) + 1 \cdot p = p$$

The variance of a Bernoulli(p) random variable is

$$\text{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

A concise way to describe the distribution is $\mathbb{P}(X = x) = (1-p)^{1-x}p^x$ for $x \in \{0,1\}$.

Definition 12. Let $A \subseteq \Omega$ be an event. We define the **indicator** of A, $\mathbb{1}\{A\}$ to be the random variable

$$\mathbb{1}\{A\}(\omega) = \begin{cases} 0 & \omega \notin A, \\ 1 & \omega \in A. \end{cases}$$

We note that $\mathbb{1}\{A\}$ follows the Bernoulli(p) distribution where $p = \mathbb{P}(A)$.

The expectation of the indicator random variable is

$$\mathbb{E}[\mathbb{1}\{A\}] = \mathbb{P}(A)$$

The variance of the indicator random variable is

$$\text{var}[\mathbb{1}\{A\}] = \mathbb{P}(A)(1 - \mathbb{P}(A))$$

An important property of indicator random variables (and Bernoulli random variables) is that $X = X^k$ for any positive integer k, since X can only take on values in the set $\{0,1\}$ and both $0^k = 0, 1^k = 1$. Therefore, $X(\omega) = X^k(\omega)$ for all outcomes $\omega \in \Omega$.

The following is a method for computing the **variance of dependent indicators** or the variance of random variables which can be written as the sum of indicators which are not independent:

Let X be written as the sum of n identically distributed indicators, where n is a positive integer, and the indicators are not assumed to be independent, i.e.

$$X = \mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_n\}$$

We can use linearity of expectation to compute the expectation to be

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{A_i\}] = \sum_{i=1}^{n} \mathbb{P}(A_i) = n\mathbb{P}(A_1)$$

Next, we compute $\mathbb{E}[X^2] = \mathbb{E}[(\mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_n\})^2]$. We note that the square has two types of terms.

There are n like-terms such as $\mathbb{1}\{A_1\}^2$ which appear in the following form:

$$\sum_{i=1}^{n} \mathbb{1}\{A_1\}^2 = \sum_{i=1}^{n} \mathbb{1}\{A_i\} = \mathbb{1}\{A_1\} + \cdots + \mathbb{1}\{A_n\} = X$$

There are also $n(n-1)$ cross-terms such as $\mathbb{1}\{A_1\}\mathbb{1}\{A_2\}$ which appear in the following form:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}\{A_i\}\mathbb{1}\{A_j\} = \mathbb{1}\{A_1\}\mathbb{1}\{A_2\} + \cdots + \mathbb{1}\{A_{n-1}\}\mathbb{1}\{A_n\}$$

Observing the cross-terms more closely, we consider the term $\mathbb{1}\{A_i\}\mathbb{1}\{A_j\}$ and notice that it is the product of two indicators, giving us another indicator since the product of either 0 or 1 is also 0 or 1. In particular, the product is 1 if and only if each indicator is 1, i.e.

$$\mathbb{P}(\mathbb{1}\{A_i\}\mathbb{1}\{A_j\} = 1) = \mathbb{P}(A_i \cap A_j)$$

Therefore, we can re-write the sum as follows:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}\{A_i\}\mathbb{1}\{A_j\} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}\{A_i \cap A_j\}$$

From the above, we thus obtain

$$X^2 = X + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}\{A_i \cap A_j\}$$

The expectation of the square is hence

$$\mathbb{E}[X^2] = \mathbb{E}[X] + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}[\mathbb{1}\{A_i \cap A_j\}] = n\mathbb{P}(A_i) + n(n-1)\mathbb{P}(A_i \cap A_j)$$

Finally, we obtain the variance to be

$$\text{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = n\mathbb{P}(A_1) + n(n-1)\mathbb{P}(A_1 \cap A_2) - n^2\mathbb{P}(A_1)^2$$

Definition 13. The **binomial distribution with parameters** n **and** p where n is a positive integer and $p \in [0, 1]$, denoted as Binomial(n, p), describes the number of successes when we conduct n independent trials, where each trial has a probability p of success.

The binomial distribution is found by the argument that the probability of having a series of trials with k successes (and hence $n-k$ failures) is $p^k(1-p)^{n-k}$ multiplied by the number of ways to achieve k successes in n trials or $\binom{n}{k}$, i.e.

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, \ldots, n\}$$

We can calculate the expectation of the binomial distribution by making a connection with the Bernoulli distribution. Let X_i be the indicator variable for the event that trial i is a success, for $i = 1, \ldots, n$. Then $X = X_1 + \cdots + X_n$ where each indicator variable X_i is 1 or 0 depending on whether or not the trial i is a success, so taking the sum of all the indicator variables gives us the total number of successes in all n trials.

The expectation of the binomial distribution is

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X_1 + \cdots + X_n] \\
&= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \\
&= p + \cdots + p \\
&= np
\end{aligned}$$

The variance of the binomial distribution is

$$\begin{aligned}
\text{var}[X] &= \text{var}(X_1 + \cdots + X_n) \\
&= \text{var}[X_1] + \cdots + \text{var}[X_n] \\
&= p(1-p) + \cdots + p(1-p) \\
&= np(1-p)
\end{aligned}$$

Definition 14. The random variables $X_1, \ldots, X_n$ are an example of **independent and identically distributed** or **i.i.d.** random variables.

Since each trial is independent of each other, the variables $X_1, \ldots, X_n$ are independent.

Definition 15. The **geometric distribution with parameter** p where $p \in (0, 1)$, denoted by Geometric(p), describes the number of trials required to obtain the first success, assuming that each trial is independent and has a probability of success p.

If it takes exactly x trials to obtain the first success, there were first $x-1$ failures, each with probability $1-p$, followed by one success, with probability p, giving us the distribution

$$\mathbb{P}(X = x) = (1-p)^{x-1}p, \quad x \text{ is a positive integer}$$

We can solve for the expectation of the geometric distribution with a recursive relation since with probability p, we obtain a success and are done, and with probability $1-p$, we obtain a failure and are back where we started.

The expectation of the geometric distribution is

$$\begin{aligned}\mathbb{E}[X] &= p \cdot 1 + (1-p) \cdot (1 + \mathbb{E}[X]) \\ &= \frac{1}{p}\end{aligned}$$

We can first compute

$$\mathbb{E}[X^2] = p \sum_{x=1}^{\infty} x^2 (1-p)^{x-1} = p \cdot \frac{2-p}{p^3} = \frac{2-p}{p^2}$$

The variance of the geometric distribution is thus

$$\text{var}[X] = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

Theorem 6. (Minimum of Geometric RVs) Let $X \sim \text{Geometric}(p)$ and $Y \sim \text{Geometric}(q)$ be independent random variables. Then

$$\min\{X, Y\} \sim \text{Geometric}(p + q - pq)$$

Proof. We can use the tail probabilities to simplify the derivation. If the minimum of X and Y is greater than z, then both X and Y are greater than z, i.e.

$$\begin{aligned}\mathbb{P}(\min\{X, Y\} > z) &= \mathbb{P}(X > z, Y > z) \\ &= \mathbb{P}(X > z)\mathbb{P}(Y > z) \\ &= (1-p)^z(1-q)^z \\ &= (1 - (p + q - pq))^z\end{aligned}$$

This gives us the tail probability of a geometric random variable with parameter $p+q-pq$. □

In the Coupon's Collector problem, there are n different coupons that you would like to collect, where n is a positive integer. Every time you buy an item from the store, you receive a random coupon, where each of the n coupons is equally likely to appear. We would like to find the expected number of items we must buy before we collect every coupon.

Let T_i be the number of items it takes to collect the i^{th} coupon, for $i = 1, \ldots, n$. In other words, starting from when you have seen $i-1$ distinct coupons, T_i represents the additional number of items you must purchase before you see a coupon you have not seen before. Therefore, the time to collect all coupons is $T = \sum_{i=1}^{n} T_i$.

Once we have collected $i-1$ coupons, there are $n-i+1$ coupons we have not seen yet, so the probability that the next item we buy comes with a coupon we have not seen is $(n-i+1)/n$. Regarding each object bought as an independent trial, we see that $T_i \sim \text{Geometric}(p)$, where $p = (n-i+1)/n$. By linearity of expectation,

$$\mathbb{E}[T] = \sum_{i=1}^{n} \mathbb{E}[T_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{i=1}^{n} \frac{1}{i} = nH_n$$

where H_n is the n^{th} harmonic sum, $H_n = \sum_{i=1}^{n} \frac{1}{i}$.

A good approximation to H_n is the $\ln n + \gamma$, where γ is the **Euler-Mascheroni constant**,

$$\gamma = \lim_{n \to \infty} (H_n - \ln n) \approx 0.577$$

Definition 16. An important property of the geometric distribution is that it is **memoryless**, which means that a random variable following the geometric distribution depends only on its current state and not the past.

Theorem 7. (Memoryless Property) The geometric distribution satisfies, for all $s, t \in \mathbb{N}$ with $s < t$,

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$$

Proof. We can use the rules of conditional probability to observe the following:

$$\begin{aligned}
\mathbb{P}(X > s + t \mid X > s) &= \frac{\mathbb{P}(X > s + t, X > s)}{\mathbb{P}(X > s)} \\
&= \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} \\
&= \frac{(1-p)^{s+t}}{(1-p)^{s}} \\
&= (1-p)^{t} = \mathbb{P}(X > t)
\end{aligned}$$

□

We can consider a slight generalization of the geometric distribution, where we would like to find the number of trials we need until we obtain k successes if we have independent trials, each with probability of success p. The probability that we require x trials is equivalent to the probability that we have k successes and $x-k$ failures, i.e. $p^k(1-p)^{x-k}$, where the last success must occur on the x^{th} trial, giving us $\binom{x-1}{k-1}$ ways to distribute the remaining successes.

Definition 17. The **negative binomial distribution with parameter** p **of order** k, where $p \in [0,1]$ and k is a positive integer, is

$$P(X=x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots$$

When $k=1$, we simply have the geometric distribution.

To compute the expectation and variance, we can use linearity of expectation where we let X_i be the number of trials it takes to obtain the i^{th} success, starting after we have already observed $i-1$ successes, for $i = 1, \ldots, n$. Then, each $X_i \sim \text{Geometric}(p)$ and $X = \sum_{i=1}^{k} X_i$.

The expectation of the negative binomial distribution is

$$\mathbb{E}[X] = \sum_{i=1}^{k} \mathbb{E}[X_i] = \frac{k}{p}$$

The variance of the negative binomial distribution is

$$\text{var}[X] = \sum_{i=1}^{k} \text{var}[X_i] = \frac{k(1-p)}{p^2}$$

Definition 18. The **Poisson distribution with parameter** λ where $\lambda > 0$, denoted by Poisson(λ), can be viewed as an approximation to the binomial distribution where we let the number of trials $n \to \infty$ and the probability of success $p \to 0$ such that the mean $\mathbb{E}[X] = np$ remains a fixed value λ.

These assumptions tell us when the approximation is reasonable, i.e. the probability of success should be low and the number of trials should be high such that the product np is roughly between 1 and 10.

The probability distribution is

$$\begin{aligned}\mathbb{P}(X=x) &= \binom{n}{x}p^x(1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+2)(n-x+1)}{x!}p^x(1-p)^{n-x} \\ &\approx \frac{n^x p^x}{x!}(1-\frac{\lambda}{n})^n\end{aligned}$$

We thus have the Poisson distribution

$$\mathbb{P}(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathbb{N}$$

The expectation of the Poisson distribution is

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x\frac{\lambda^x e^{-\lambda}}{x!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda$$

To compute the variance, we first calculate $\mathbb{E}[X(X-1)]$ to be the following:

$$\mathbb{E}[X(X-1)] = \sum_{x=2}^{\infty} x(x-1)\frac{\lambda^x e^{-\lambda}}{x!} = \lambda^2 e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} = \lambda^2 e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = \lambda^2$$

By linearity of expectation, the variance is thus

$$\text{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Theorem 8. (Poisson Merging) Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent random variables. Then

$$X + Y \sim \text{Poisson}(\lambda + \mu)$$

Proof. We can show this by computing the distribution of $X + Y$:

$$\begin{aligned}\mathbb{P}(X+Y=z) &= \sum_{j=0}^{z}\mathbb{P}(X=j, Y=z-j) = \sum_{j=0}^{z}\frac{\lambda^j e^{-\lambda}}{j!}\frac{\mu^{z-j}e^{-\mu}}{(z-j)!} \\ &= \frac{e^{-(\lambda+\mu)}}{z!}\sum_{j=0}^{z}\frac{z!}{j!(z-j)!}\lambda^j\mu^{z-j} = \frac{e^{-(\lambda+\mu)}}{z!}\sum_{j=0}^{z}\binom{z}{j}\lambda^j\mu^{z-j} \\ &= \frac{(\lambda+\mu)^z e^{-(\lambda+\mu)}}{z!}\end{aligned}$$

We observe this to be the Poisson distribution with parameter $\lambda + \mu$. □

Theorem 9. (Poisson Splitting) Suppose that $X \sim \text{Poisson}(\lambda)$ and that conditioned on $X = x$, Y follows the Binomial(x, p) distribution and Z follows the Binomial$(X, 1-p)$ distribution, such that $Y + Z = X$. Then $Y \sim \text{Poisson}(\lambda p)$, $Z \sim \text{Poisson}(\lambda(1-p))$, and Y and Z are independent.

Proof. We can use the definition of conditional probability and show that Y has the correct distribution:

$$\begin{aligned}
\mathbb{P}(Y = y) &= \sum_{x=y}^{\infty} \mathbb{P}(X = x, Y = y) = \sum_{x=y}^{\infty} \mathbb{P}(X = x)\mathbb{P}(Y = y \mid X = x) \\
&= \sum_{x=y}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \binom{x}{y} p^y (1-p)^{x-y} = e^{-y} \sum_{x=y}^{\infty} \frac{\lambda^x}{x!} \frac{x!}{y!(x-y)!} p^y (1-p)^{x-y} \\
&= \frac{(\lambda p)^y e^{-\lambda}}{y!} \sum_{x=y}^{\infty} \frac{(\lambda(1-p))^{x-y}}{(x-y)!} = \frac{(\lambda p)^y e^{-\lambda}}{y} e^{\lambda(1-p)} \\
&= \frac{(\lambda p)^y e^{-\lambda p}}{y!}
\end{aligned}$$

We observe this to be the Poisson distribution with parameter λp. Likewise, we can replace p with $1-p$ to find that Z is the Poisson distribution with parameter $\lambda(1-p)$. □

2.3.5 Inequalities

Often, probability distributions can be difficult to compute exactly, so the following are several important bounds.

Theorem 10. (Markov's Inequality) Let X be a random variable, f be an increasing, non-negative function, and $a \in \mathbb{R}$ such that $f(a) \neq 0$. Then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[f(X)]}{f(a)}$$

(Weak Markov's Inequality) For non-negative random variable X and $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. Let $\mathbb{1}\{X \geq a\}$ be the indicator that $X \geq a$ and define $h(X) = f(a)\mathbb{1}\{X \geq a\}$. We claim $h(X) \geq f(X)$ always:

1. If $X < a$, then $\mathbb{1}\{X \geq a\} = 0$, so $h(X) = 0 \geq f(X)$ (since f is non-negative).
2. If $X \geq a$, then $h(X) = f(a) \geq f(X)$ (since f is increasing).

Then we have

$$\mathbb{E}[f(X)] \geq \mathbb{E}[h(X)] = \mathbb{E}[f(a)\mathbb{1}\{X \geq a\}] = f(a)\mathbb{E}[\mathbb{1}\{X \geq a\}] = f(a)\mathbb{P}(X \geq a)$$

□

Theorem 11. (Chebyshev's Inequality) Let X be a random variable and $a > 0$. Then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\text{var}[X]}{a^2}$$

Proof. Let $Y = |X - \mathbb{E}[X]|$ and $f(y) = y^2$. Since f is an increasing function, apply Markov's inequality:

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y^2]}{a^2}$$

where $\mathbb{E}[Y^2] = \mathbb{E}[|X - \mathbb{E}[X]|^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{var}[X]$ □

Theorem 12. (Cauchy-Schwarz Inequality) Provided that $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$, we have that

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

2.3.6 Weak Law of Large Numbers

We can now justify why the expectation is called the long-run average of a sequence of values.

Suppose that $X_1, X_2, X_3, \ldots$ are i.i.d. random variables, which we can think of as successive measurements of a true variable X. The idea is that X is some quantity which we wish to measure, and X follows some probability distribution with unknown parameters: mean μ and variance σ^2. For each positive integer i, the random variable X_i is a measurement of X. In particular, this means that X_i also has mean μ and variance σ^2. We are interested in the average of the samples we collect,

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

Using linearity of expectation, we verify that

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = \frac{1}{n}(n\mu) = \mu$$

We therefore call $\bar{X}_n$ an **unbiased estimator** of μ.

We compute the variance of $\bar{X}_n$ to be

$$\text{var}[\bar{X}_n] = \frac{1}{n^2}(\text{var}[X_1] + \cdots + \text{var}[X_n]) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Theorem 13. (Weak Law of Large Numbers) For all $\varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \to 0$$

Proof. Using Chebyshev's inequality, we find that

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\text{var}[\bar{X}_n]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

which tends to 0 as $n \to \infty$. □

The Weak Law of Large Numbers states that increasing the number of samples decreases the probability that the sample average will be far from the true average. When such a case holds, we say that $\bar{X}_n$ **converges in probability** to μ as $n \to \infty$.

2.3.7 Chernoff Bounds

We consider what happens when we apply Markov's inequality to the increasing non-negative function $f(x) = e^{\theta x}$ (for $\theta > 0$).

Definition 19. The **Chernoff bound** states that

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[e^{\theta X}]}{e^{\theta X}}$$

where we optimize over the values of θ in search of the best possible bound.

2.4 Regression and Conditional Expectation

2.4.1 Covariance

A fundamental question in statistics is the matter of estimating the value of Y as a function of X given a set of data points $\{(X_i, Y_i)\}_{i=1}^n$.

Definition 1. The **covariance** of two random variables X and Y is defined as

$$\text{cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

or the product of the deviations of the two variables from their respective means.

If the covariance is positive, then we say the variables are **positively correlated**, which means that X and Y tend to fluctuate in the same direction. If the covariance is negative, then we say that the variables are **negatively correlated**, which means that X and Y tend to fluctuate in opposite directions.

Theorem 1. Let X and Y be random variables. Then

$$\text{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Proof. We obtain this from the definition of covariance:

$$\begin{aligned}\operatorname{cov}[X,Y] &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

□

In addition, if X and Y are independent random variables, then

$$\operatorname{cov}[X,Y] = 0$$

The following is an example of a case where two variables may have a covariance of 0 but are not independent:

Suppose we pick a point uniformly randomly from $\{(1,0),(0,1),(-1,0),(0,-1)\}$ where X is the x-coordinate and Y is the y-coordinate of the point. Then $\mathbb{E}[XY] = 0$ and $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, so $\operatorname{cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$, but

$$0 = \mathbb{P}(X=0, Y=0) \neq \mathbb{P}(X=0)\mathbb{P}(Y=0) = 1/4$$

The covariance and variance are also related. Let X be a random variable. Then

$$\operatorname{var}[X] = \operatorname{cov}[X,X]$$

since $\operatorname{cov}[X,X] = \mathbb{E}[X^2] - \mathbb{E}[X]\mathbb{E}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \operatorname{var}[X]$.

Finally, the following holds for any random variables X and Y:

$$\operatorname{var}[X+Y] = \operatorname{var}[X] + \operatorname{var}[Y] + 2 \cdot \operatorname{cov}[X,Y]$$

Theorem 2. Suppose X_1, X_2, Y_1, Y_2 are random variables and $a \in \mathbb{R}$ is a constant. Then the following properties hold:

$$\begin{aligned}\operatorname{cov}[X,Y] &= \operatorname{cov}[Y,X] \\ \operatorname{cov}[X_1+X_2,Y] &= \operatorname{cov}[X_1,Y] + \operatorname{cov}[X_2,Y] \\ \operatorname{cov}[X,Y_1+Y_2] &= \operatorname{cov}[X,Y_1] + \operatorname{cov}[X,Y_2] \\ a\operatorname{cov}[X,Y] &= \operatorname{cov}[aX,Y] = \operatorname{cov}[X,aY]\end{aligned}$$

i.e. the covariance is symmetric and bilinear (linear in each of its arguments).

Theorem 3. Let X be a non-constant random variable. Then the **standard form** of X is

$$X^* = \frac{X - \mathbb{E}[X]}{\sigma_X}$$

i.e. zero mean ($\mathbb{E}[X^*] = 0$) and unit variance ($\text{var}[X^*] = \mathbb{E}[(X^*)^2] = 1$).

Definition 2. The **correlation** or **Pearson's correlation coefficient** of two random variables is

$$\rho = \text{corr}[X, Y] = \frac{\text{cov}[X, Y]}{\sigma_X \sigma_Y}$$

Theorem 4. The correlation of two random variables X and Y is

$$\text{corr}[X, Y] = \text{cov}[X^*, Y^*] = \mathbb{E}[X^* Y^*]$$

Proof. We can calculate the covariance of X^* and Y^* using the properties of covariance:

$$\text{cov}[X^*, Y^*] = \text{cov}\Big[\frac{X - \mathbb{E}[X]}{\sigma_X}, \frac{Y - \mathbb{E}[Y]}{\sigma_Y}\Big] = \frac{\text{cov}[X, Y]}{\sigma_X \sigma_Y}$$

□

Theorem 5. If X and Y are non-constant random variables, then

$$-1 \leq \text{corr}[X, Y] \leq 1$$

with equality if and only if Y is a linear function of X.

Correlation is therefore a standardized version of the covariance and a useful measure of the degree of linear dependence between two variables X And Y.

2.4.2 LLSE

Definition 3. Let X and Y be random variables. The **least linear squares estimate (LLSE)** of Y given X is defined as

$$L(Y \mid X) = \mathbb{E}[Y] + \frac{\text{cov}[X, Y]}{\text{var}[X]}(X - \mathbb{E}[X])$$

We note that LLSE is a random variable, in particular a function of X.

Theorem 6. The orthogonality property states that the LLSE satisfies

$$\mathbb{E}[Y - L(Y \mid X)] = 0$$
$$\mathbb{E}[(Y - L(Y \mid X))X] = 0$$

Theorem 7. Define $\mathcal{L}(X) = \{aX + b : a, b \in R\}$ to be the set of linear functions of X. Then for any $aX + b \in \mathcal{L}(X)$,

$$\mathbb{E}[(Y - L(Y \mid X))^2] \leq \mathbb{E}[(Y - aX - b)^2]$$

i.e. when we estimate Y, $L(Y \mid X)$ has the lowest mean squared error of any linear function of X.

Proof. We first use the orthogonality property, from which it follows that

$$\mathbb{E}[(Y - L(Y \mid X))(aX + b)] = 0$$

for any $aX + b \in \mathcal{L}(X)$. Let $\hat{Y} = L(Y \mid X)$. Then

$$\begin{aligned}
\mathbb{E}[(Y - aX - b)^2] &= \mathbb{E}[((Y - \hat{Y}) + (\hat{Y} - aX - b))^2] \\
&= \mathbb{E}[(Y - \hat{Y})^2] + 2\mathbb{E}[(Y - \hat{Y})(\hat{Y} - aX - b)] + \mathbb{E}[(\hat{Y} - aX - b)^2] \\
&= \mathbb{E}[(Y - \hat{Y})^2] + \mathbb{E}[(\hat{Y} - aX - b)^2]
\end{aligned}$$

In the second line, the term $\mathbb{E}[(Y - \hat{Y})(\hat{Y} - aX - b)]$ vanishes by the orthogonality property because $L(Y \mid X) - aX - b \in \mathcal{L}(X)$. We wish to minimize the quantity above, which represents the mean squared error of $aX + b$ as a predictor of Y and since the expectation of a non-negative random variable is always non-negative, the mean squared error is minimized when $aX + b = L(Y \mid X)$. □

2.4.3 Conditional Expectation

Definition 4. If $\mathbb{P}(Y = y) > 0$, we define the **conditional expectation**, or the expectation of X with respect to the probability distribution of X conditioned on $Y = y$ as

$$\mathbb{E}[X \mid Y = y] = \sum_x x\mathbb{P}(X = x \mid Y = y)$$

For every possible value of Y, $\mathbb{E}[X \mid Y = y]$ is a real number.

Definition 5. Let X and Y be random variables. Then $\mathbb{E}(X \mid Y)$ is also a random variable, called the **conditional expectation of X given Y**, which has the value $\mathbb{E}[X \mid Y = y]$ with probability $\mathbb{P}(Y = y)$.

We note that $\mathbb{E}(X \mid Y)$ is a function of Y and hence a random variable.

Theorem 8. (Law of Iterated Expectation) Let X and Y be random variables. Then

$$\mathbb{E}[\mathbb{E}(X \mid Y)] = \mathbb{E}[X]$$

Proof. We can compute $\mathbb{E}[\mathbb{E}(X \mid Y)]$ with respect to the probability distribution of Y as follows:

$$\begin{aligned}\mathbb{E}[\mathbb{E}(X \mid Y)] &= \sum_y \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y) = \sum_y \Big[\sum_x x\mathbb{P}(X = x \mid Y = y)\Big]\mathbb{P}(Y = y) \\ &= \sum_y \sum_x x\mathbb{P}(X = x, Y = y) = \sum_x x\Big[\sum_y \mathbb{P}(X = x, Y = y)\Big] \\ &= \sum_x x\mathbb{P}(X = x) = \mathbb{E}[X]\end{aligned}$$

□

2.4.4 MMSE

Theorem 9. Let X and Y be random variables and let $\phi(X)$ be any function of X. Then

$$\mathbb{E}[(Y - \mathbb{E}(Y \mid X))]\phi(X)] = 0$$

Proof. We condition on X and obtain

$$\begin{aligned}\mathbb{E}[(Y - \mathbb{E}(Y \mid X))\phi(X) \mid X] &= \phi(X)\mathbb{E}(Y - \mathbb{E}(Y \mid X) \mid X) \\ &= \phi(X)(\mathbb{E}(Y \mid X) - \mathbb{E}(\mathbb{E}(Y \mid X) \mid X)) \\ &= \phi(X)(\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X)) = 0\end{aligned}$$

Therefore, by the law of iterated expectation, $\mathbb{E}[(Y - \mathbb{E}(Y \mid X))]\phi(X)] = 0$. □

Definition 6. The **minimum mean square error (MMSE)** estimator of Y given X is the random variable $f(X)$ which minimizes the mean squared error, i.e. for any function g,

$$\mathbb{E}[(Y - f(X))^2] \leq \mathbb{E}[(Y - g(X))^2]$$

Theorem 10. Let X and Y be random variables. Then the MMSE of Y given X is $\mathbb{E}(Y \mid X)$, i.e. for any function g,

$$\mathbb{E}[(Y - \mathbb{E}(Y \mid X))^2] \leq \mathbb{E}[(Y - g(X))^2]$$

Proof. Let $\hat{Y} = \mathbb{E}(Y \mid X)$. By the orthogonality property,

$$\begin{aligned}\mathbb{E}[(Y - g(X))^2] &= \mathbb{E}[(Y - \hat{Y} + \hat{Y} - g(X))^2] \\ &= \mathbb{E}[(Y - \hat{Y})^2] + 2\mathbb{E}[(Y - \hat{Y})(\hat{Y} - g(X))] + \mathbb{E}[(\hat{Y} - g(X))^2] \\ &= \mathbb{E}[(Y - \hat{Y})^2] + \mathbb{E}[(\hat{Y} - g(X))^2]\end{aligned}$$

□

2.4.5 Conditional Variance

Definition 7. Let X and Y be random variables. we define $\text{var}[X \mid Y = y]$ to be the variance of the conditional probability distribution $\mathbb{P}(X = x \mid Y = y)$. Furthermore, the **conditional variance** $\text{var}(X \mid Y)$ is defined to be the random variable that takes on the value $\text{var}[X \mid Y = y]$ with probability $\mathbb{P}(Y = y)$.

We note that $\text{var}(X, Y)$ is a function of Y, analogously to $\mathbb{E}(X \mid Y)$.

Theorem 11. (Law of Total Variance) Let X And Y be random variables. Then

$$\text{var}[X] = \mathbb{E}[\text{var}(X \mid Y)] + \text{var}[\mathbb{E}(X \mid Y)]$$

Proof. We calculate each term of the variance by the law of iterated expectation:

$$\begin{aligned}
\text{var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}[\mathbb{E}(X^2 \mid Y)] - \mathbb{E}[\mathbb{E}(X \mid Y)]^2 \\
&= \mathbb{E}[\mathbb{E}(X^2 \mid Y)] - \mathbb{E}[E(X \mid Y)^2] + \mathbb{E}[E(X \mid Y)^2] - \mathbb{E}[\mathbb{E}(X \mid Y)]^2 \\
&= \mathbb{E}[\mathbb{E}(X^2 \mid Y) - \mathbb{E}(X \mid Y)^2] + \text{var}[\mathbb{E}(X \mid Y)] \\
&= \mathbb{E}[\text{var}(X \mid Y)] + \text{var}[\mathbb{E}(X \mid Y)]
\end{aligned}$$

□

2.5 Continuous Probability

2.5.1 Continuous Probability

Often, we would like to study random variables that taken on a continuous range of values, an uncountable number of values.

Definition 1. The **density function** or **probability density function (PDF)** of a continuous random variable X is a real-valued function f_X such that

1. $f_X(x) \geq 0, \forall x \in \mathbb{R}$, i.e. f_X is non-negative
2. $\int_{\mathbb{R}} f_X(x)dx = 0$, i.e. f_X is normalized

We can define the probability that X lies in some interval $[a, b]$ to be

$$\mathbb{P}(X \in [a, b]) = \int_a^b f_X(x)dx$$

The probability of an interval in $\mathbb{R}$ is interpreted as the area under the density function above the interval.

Definition 2. The **cumulative distribution function (CDF)** of X is defined as

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(x')dx'$$

To obtain the PDF from the CDF, we can use

$$f_X(x) = \frac{d}{dx}F_X(x)$$

The following is an interpretation for the density function:

$$\mathbb{P}(X < x + dx) - \mathbb{P}(X < x) = \mathbb{P}(X \in (x, x + dx)) = f_X(x)dx$$

2.5.2 Continuous Analogues of Discrete Results

Definition 3. The **expectation** of a continuous random variable X is

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x$$

Similarly, the expectation of a function is

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x$$

Definition 4. The **variance** of a continuous random variable remains

$$\mathrm{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Definition 5. The **joint density** of two random variables X and Y is $f_{X,Y}$ and represents everything there is to know about the two random variables. The joint distribution must satisfy the normalization condition

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$$

Definition 6. We say that X and Y are **independent** if and only if the joint density factors as such:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

Definition 7. To obtain the **marginal distribution** of X from the joint distribution, we integrate out the unnecessary variables, i.e.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}y$$

The joint density can be extended to multiple random variables $X_1, \ldots, X_n$, where n is any positive integer, satisfying the normalization condition

$$\int_{\mathbb{R}^n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$$

We can use the joint density to compute probabilities. Consider a region $J \subseteq \mathbb{R}^2$. The probability that $(X, Y) \in J$ is

$$\mathbb{P}((X, Y) \in J) = \int_J f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

To calculate the expectation of a function of multiple random variables $X_1, \ldots, X_n$, we compute

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

Theorem 1. (Continuous Tail Sum Formula) Let X be a non-negative random variable. Then

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\,\mathrm{d}x$$

2.5.3 Important Continuous Distributions

Definition 8. In the **Uniform**$[0, 1]$ **distribution**, X is chosen uniformly randomly from the interval $[0, 1]$. The density function is

$$f_X(x) = 1, \quad 0 < x < 1$$

The CDF is found by integrating:

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x \leq 1 \\ 1, & x > 1 \end{cases}$$

To compute the expectation, we simply observe that the distribution is symmetric about $x = \frac{1}{2}$ and obtain that

$$\mathbb{E}[X] = \frac{1}{2}$$

To compute the variance, we find that $\mathbb{E}[X^2] = \int_0^1 x^2\,\mathrm{d}x = \frac{1}{3}$ and thus

$$\mathrm{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Suppose that X and Y are i.i.d. Uniform$[0,1]$ random variables. Since they are independent, the joint distribution is simply the product of their respective density functions:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = 1, \quad 0 < x < 1, 0 < y < 1$$

Definition 9. In the **Uniform**$[a,b]$ **distribution**, X is chosen uniformly randomly in the interval $[a,b]$. The density function is

$$f_X(x) = 1, \quad a < x < b$$

The CDF is found by integrating:

$$F_X(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a < x \le b \\ 1, & x > b \end{cases}$$

We observe that if U is the uniform distribution from $[0,1]$, then $a + (b-a)U$ has the same CDF as X. We can use this finding to compute the expectation and variance.

We compute the expectation to be

$$\mathbb{E}[X] = \mathbb{E}[a + (b-a)U] = a + \frac{b-a}{2} = \frac{a+b}{2}$$

We compute the variance to be

$$\text{var}[X] = \text{var}[a + (b-a)U] = \frac{(b-a)^2}{12}$$

Definition 10. The **exponential distribution with parameter** λ, where $\lambda > 0$, is given by the density function

$$f_T(t) = \lambda e^{-\lambda t}, \quad t > 0$$

The CDF is found by integrating:

$$F_X(x) = \begin{cases} 0, & t < 0 \\ 1 - e^{-\lambda t}, & t \ge 0 \end{cases}$$

To compute the expectation, we use integration by parts to obtain that

$$\mathbb{E}[T] = \int_0^\infty t\lambda e^{-\lambda t}\,\mathrm{d}t = \frac{1}{\lambda}$$

To compute the variance, we find that $\mathbb{E}[T^2] = \int_0^\infty t^2\lambda e^{-\lambda t}\,\mathrm{d}t = \frac{2}{\lambda^2}$ and thus

$$\text{var}[T] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Theorem 2. Let $T > 0$ be a random variable with a continuous CDF. Then T satisfies the **memoryless property**, i.e. for all $s, t > 0$,

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t)$$

if and only if $T \sim \text{Exponential}(\lambda)$ for some $\lambda > 0$.

Theorem 3. Let n be a positive integer and $T_1, \ldots, T_n$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n > 0$ respectively. Then the minimum of the random variables is also exponentially distributed:

$$\min\{T_1, \ldots, T_n\} \sim \text{Exponential}(\lambda_1 + \cdots + \lambda_n)$$

Proof. We consider the tail probabilities and find that:

$$\begin{aligned}
\mathbb{P}(\min\{T_1, \ldots, T_n\} > t) &= \mathbb{P}(T_1 > t, \ldots, T_n > t) \\
&= \mathbb{P}(T_1 > t) \cdots \mathbb{P}(T_n > t) \\
&= e^{-\lambda_1 t} \cdots e^{-\lambda_n t} \\
&= e^{-(\lambda_1 + \cdots + \lambda_n)t}
\end{aligned}$$

We observe that this is an exponential distribution with parameter $\lambda_1 + \cdots + \lambda_n$. □

2.5.4 Conditional Probability

Definition 11. The **law of total probability** states that if we have an event A and would like to calculate $\mathbb{P}(A)$, then we can integrate over the density as follows:

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A \mid X = x) f_X(x) \, \mathrm{d}x$$

Definition 12. The **conditional density** of Y given $X = x$ is

$$F_{Y|X}(y \mid x) = \mathbb{P}(Y \leq y \mid X = x) = \int_{-\infty}^{y} f_{Y|X}(y' \mid x) \, \mathrm{d}y'$$

We can compute the probability of an event $[a, b]$, where $a < b$, to be

$$\mathbb{P}(Y \in [a, b] \mid X = x) = F_{Y|X}(b \mid x) - F_{Y|X}(a \mid x) = \int_a^b f_{Y|X}(y \mid x) \, \mathrm{d}y$$

2.5.5 Functions of Random Variables

Theorem 4. (Change of Variables) Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$ be differentiable and one-to-one. Let h be the inverse of g. Then

$$f_Y(y) = f_X(h(y))|h'(y)|$$

Proof. We can show this by first manipulating the CDF, then differentiating. Considering the case where g is strictly increasing,

$$F_Y(y) = \mathbb{P}(Y < y) = \mathbb{P}(g(X) < y) = \mathbb{P}(X < h(y)) = F_X(h(y))$$

Differentiating both sides, we obtain

$$f_Y(y) = f_X(h(y))h'(y)$$

Since h is strictly increasing, $h'(y) > 0$ and the change of variables equation holds. Now, considering the case where g is strictly decreasing,

$$F_Y(y) = \mathbb{P}(X > h(y)) = 1 - \mathbb{P}(X < h(y)) = 1 - F_X(h(y))$$

Differentiating both sides, we obtain

$$f_Y(y) = -f_X(h(y))h'(y)$$

Since h is strictly decreasing, $h'(y) < 0$ and the change of variables equation still holds. □

To compute the density of $Z = X + Y$, we note that $X + Y = Z \in (z, z + \mathrm{d}z)$ is equivalent to the event that X ranges over all values and $Y \in (z - x, z - x + dz)$. Therefore, we can find $\mathbb{P}(Z \in (z, z + \mathrm{d}z))$ by integrating over the joint density of X and Y to obtain

$$f_Z(z)\,\mathrm{d}z = \int_{-\infty}^{\infty} \int_{z-x}^{z-x+dz} f_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x$$

Since $\mathrm{d}z$ is an infinitesimal length, we can assume that $f_{X,Y}$ is effectively constant over the inner integral and therefore, we have

$$f_Z(z)\,\mathrm{d}z = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\,\mathrm{d}z\,\mathrm{d}x$$

Definition 13. We end up with what is known as the **convolution formula**,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,\mathrm{d}x \end{aligned}$$

To compute the density of $Z = X/Y$, we observe that $X/Y = Z \in (z, z + \mathrm{d}z)$ is equivalent to $X \in (yz, yz + y\,\mathrm{d}z)$ when $y > 0$ and $X \in (yz + y\,\mathrm{d}z, yz)$ when $y < 0$. Therefore,

$$f_Z(z) = \int_{-\infty}^{0} \int_{yz+y\,\mathrm{d}z}^{yz} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y + \int_{0}^{\infty} \int_{yz}^{yz+y\,\mathrm{d}z} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Again, since $\mathrm{d}z$ is an infinitesimal length, we can assume that $f_{X,Y}$ is effectively constant over the inner integral and therefore, we have

$$f_Z(z)\,\mathrm{d}z = \int_{\infty}^{0} f_{X,Y}(yz, y)(-y\,\mathrm{d}z)\,\mathrm{d}y + \int_{0}^{\infty} f_{X,Y}(yz, y)(y\,\mathrm{d}z)\,\mathrm{d}y = \int_{-\infty}^{\infty} |y| f_{X,Y}(yz, y)\,\mathrm{d}y\,\mathrm{d}z$$

We end up with

$$f_Z z = \int_{-\infty}^{\infty} |y| f_{X,Y}(yz, y)\,\mathrm{d}y$$

In the special case when X and Y are independent, we have

$$f_Z z = \int_{-\infty}^{\infty} |y| f_X(yz) f_Y(yy)\,\mathrm{d}y$$

2.5.6 Normal Distribution

Definition 14. We denote $\mathcal{N}(0, 1)$ to be the **standard normal distribution** or **Gaussian distribution**, with CDF

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x$$

To find the expectation, we observe that the density $f_X(x)$ depends only on x^2 and is thus symmetric about $x = 0$, giving us

$$\mathbb{E}[X] = 0$$

To find the variance, we calculate

$$\mathrm{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,\mathrm{d}x = 1$$

We can apply the change of variables technique to the standard normal distribution in order to obtain the general form of the normal distribution.

Definition 15. We denote $Y \sim \mathcal{N}(\mu, \sigma^2)$ to be the **normal distribution**, with PDF

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

and expectation $\mathbb{E}[Y] = \mu$ and $\mathrm{var}[Y] = \sigma^2$.

We note that any normal distribution can be obtained from the standard normal distribution with a scale by σ and a shift by μ.

Theorem 5. (Sums of Independent Gaussians) Let n be any positive integer and let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$. Then

$$X = X_1 + \cdots + X_n \sim \mathcal{N}\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\Big)$$

We note that for sums of independent Gaussians, variances add but standard deviations do not.

2.5.7 Central Limit Theorem

Suppose $\{X_i\}_{i \in \mathbb{N}}$ is a sequence of i.i.d. random variables with mean μ and finite variance σ^2. We define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

As $n \to \infty$, $\sqrt{n}(\bar{X}_n - \mu)$ **converges in distribution** to the normal distribution.

We recall that the Weak Law of Large Numbers tells us that as we increase the number of samples, the sample mean converges in probability to the expected value. The Central Limit Theorem states that the distribution of the sample mean also converges to a particular distribution, the normal distribution.

Theorem 6. (Central Limit Theorem) Let $\{X_i\}_{i \in \mathbb{N}}$ be a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then, for all $z \in \mathbb{R}$,

$$\mathbb{P}\Big(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq z\Big) \to \Phi(z)$$